METHOD ARTICLE

REVISED Prediction of mosquito species and population age structure using mid-infrared spectroscopy and supervised machine learning [version 3; peer review: 2 approved]

Mario González Jiménez [1*], Simon A. Babayan[2*], Pegah Khazaeli[1], Margaret Doyle[2], Finlay Walton [1], Elliott Reedy [1], Thomas Glew[1], Mafalda Viana [2], Lisa Ranford-Cartwright [2], Abdoulaye Niang[3], Doreen J. Siria[4], Fredros O. Okumu [2,4], Abdoulaye Diabaté[3], Heather M. Ferguson [2], Francesco Baldini [2], Klaas Wynne [1]

[1]School of Chemistry, University of Glasgow, Glasgow, G12 8QQ, UK
[2]Institute of Biodiversity Animal Health and Comparative Medicine, University of Glasgow, Glasgow, G12 8QQ, UK
[3]Department of Medical Biology and Public Health, Institut de Recherche en Science de la Santé (IRSS), Bobo-Dioulasso, Burkina Faso
[4]Environmental Health & Ecological Sciences Department, Ifakara Health Institute, Off Mlabani Passage, PO Box 53, Ifakara, Tanzania

[*] Equal contributors

v3 **First published:** 01 May 2019, **4**:76 (https://doi.org/10.12688/wellcomeopenres.15201.1)
Second version: 07 Aug 2019, **4**:76 (https://doi.org/10.12688/wellcomeopenres.15201.2)
Latest published: 16 Sep 2019, **4**:76 (https://doi.org/10.12688/wellcomeopenres.15201.3)

Abstract

Despite the global efforts made in the fight against malaria, the disease is resurging. One of the main causes is the resistance that *Anopheles* mosquitoes, vectors of the disease, have developed to insecticides. *Anopheles* must survive for at least 10 days to possibly transmit malaria. Therefore, to evaluate and improve malaria vector control interventions, it is imperative to monitor and accurately estimate the age distribution of mosquito populations as well as their population sizes. Here, we demonstrate a machine-learning based approach that uses mid-infrared spectra of mosquitoes to characterise simultaneously both age and species identity of females of the African malaria vector species *Anopheles gambiae* and *An. arabiensis*, using laboratory colonies. Mid-infrared spectroscopy-based prediction of mosquito age structures was statistically indistinguishable from true modelled distributions. The accuracy of classifying mosquitoes by species was 82.6%. The method has a negligible cost per mosquito, does not require highly trained personnel, is rapid, and so can be easily applied in both laboratory and field settings. Our results indicate this method is a promising alternative to current mosquito species and age-grading approaches, with further improvements to accuracy and expansion for use with wild mosquito vectors possible through collection of larger mid-infrared spectroscopy data sets.

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
REVISED **version 3** published 16 Sep 2019	✓ report	
REVISED **version 2** published 07 Aug 2019	? report	✓ report
version 1 published 01 May 2019	✗ report	? report

1 **Yoosook Lee**, University of California, Davis, Davis, USA

2 **Thomas S. Churcher** , Imperial College London, London, UK

Keywords
Malaria, Anopheles gambiae, Anopheles arabiensis, Vector control, Machine learning, Mid-infrared spectroscopy

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding authors: Francesco Baldini (Francesco.Baldini@glasgow.ac.uk), Klaas Wynne (klaas.wynne@glasgow.ac.uk)

Author roles: González Jiménez M: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Babayan SA**: Conceptualization, Data Curation, Formal Analysis, Methodology, Project Administration, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Khazaeli P**: Investigation, Writing – Review & Editing; **Doyle M**: Investigation, Writing – Review & Editing; **Walton F**: Investigation, Writing – Review & Editing; **Reedy E**: Investigation, Writing – Review & Editing; **Glew T**: Investigation, Writing – Review & Editing; **Viana M**: Formal Analysis, Methodology, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Ranford-Cartwright L**: Conceptualization, Resources, Writing – Review & Editing; **Niang A**: Conceptualization, Writing – Review & Editing; **Siria DJ**: Conceptualization, Writing – Review & Editing; **Okumu FO**: Conceptualization, Funding Acquisition, Writing – Review & Editing; **Diabaté A**: Conceptualization, Writing – Review & Editing; **Ferguson HM**: Conceptualization, Funding Acquisition, Project Administration, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Baldini F**: Conceptualization, Data Curation, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Wynne K**: Conceptualization, Funding Acquisition, Project Administration, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Wellcome Trust through a Intermediate Fellowship in Public Health and Tropical Medicine to FO [102350]. This work was also supported by The Engineering and Physical Sciences Research Council (EPSRC) [EP/J009733/1, EP/K034995/1, EP/N508792/1, and EP/N007417/1] and Medical Research Council (MRC) [MR/P025501/1]. FB is supported by an AXA Research Fund fellowship [14-AXA-PDOC-130] and a European Molecular Biology Organization (EMBO) Long Term fellowship [43-2014]. MV is funded under the MRC/Department for International Development Concor-dat agreement, which is part of EU EDCTP2 programme [MR/N015320/1].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: González Jiménez M, Babayan SA, Khazaeli P *et al.* **Prediction of mosquito species and population age structure using mid-infrared spectroscopy and supervised machine learning [version 3; peer review: 2 approved]** Wellcome Open Research 2019, **4**:76 (https://doi.org/10.12688/wellcomeopenres.15201.3)

First published: 01 May 2019, **4**:76 (https://doi.org/10.12688/wellcomeopenres.15201.1)

REVISED **Amendments from Version 2**

This third version includes the following amendments:

We have corrected the typos that have appeared in the text.

We have clarified how we have modelled the populations.

We have changed Figure 8 to include wavelength units as well.

And we have improved the bibliography with two new references.

We sincerely appreciate the comments made by the reviewers during this process.

Any further responses from the reviewers can be found at the end of the article

Between 2000 and 2015, insecticide-based control interventions targeting mosquito vectors averted an estimated 537 million malaria cases[1]. Nevertheless, malaria still kills hundreds of thousands of people each year (445,000 in 2016), mainly in sub-Saharan Africa[2]. Additionally, there is concern that progress may have stalled after more than a decade of success in global malaria control[2]. Of major concern is the increase in insecticide resistance among mosquito populations throughout Africa[3], which is degrading the lethality and effectiveness of vector control tools, notably indoor residual spraying (IRS) and long-lasting insecticide treated nests (LLINs) which have been the cornerstones of malaria control in the past decades[4]. Indeed, much of the effectiveness of LLINs and IRS comes from community-wide reductions in vector population size, not merely from preventing people from getting bitten[5].

Measurement of female mosquito vector survival is an important biological determinant of malaria transmission intensity[6,7]. This is because malaria parasites (*Plasmodium* spp.) require more than 10 days of incubation inside female mosquito vectors (extrinsic incubation period, EIP) before they become infectious[8–11]. While there is uncertainty about mosquito survival in the field, crude estimates suggest the median lifespan of African malaria vectors is 7–10 days[12]. Thus, only relatively old mosquitoes can transmit the parasite[13]. As a result, even minor reductions in mosquito survival can have exponential impacts on pathogen transmission[10,14]. Consequently, accurate and high-resolution estimation of both mosquito abundance and longevity is essential for the assessment of the impact of various vector control measures.

Despite the crucial importance of mosquito demography to vector control, there are few reliable tools for rapid, high-throughput monitoring of mosquito survival in the wild. Conventionally, mosquito age has been approximated by classifying females (the only sex that transmits malaria) into groups based on their reproductive status as assessed through observation of their ovarian tracheoles[15]. This widely-employed technique distinguishes females who have not yet laid eggs (nulliparous) from those that have laid at least one egg batch (parous), with the latter group assumed to be older than the former because the gonotrophic cycle between blood feeding and oviposition takes ~ 4 days. While useful for approximating general patterns of survival[16], this method is crude and cannot distinguish between females who have laid eggs only once or multiple times. Alternatively, more refined methods have been developed to estimate the number of gonotrophic cycles a female mosquito has gone through based on follicular relics or dilatations formed during each oviposition[17], although the conversion between gonotrophic cycles and actual age is imprecise (especially now that LLINs are limiting regular access to blood-meals)[18]. While an improvement on the simple parity classification method, this approach is extremely technically demanding and time-consuming[19]. Additionally, it is unsuitable for analysis of the large sample sizes necessary for estimating mosquito population structure[20].

Given these problems with ovary-based assessment[21], there has been significant investigation of alternative, molecular-based approaches to estimate mosquito age. These methods include: counting cuticle rings representing daily growth layers of the mosquito skeletal apodemes[22], chromatographic analysis of cuticular hydrocarbon chains[23], assessment of pteridines using fluorescence techniques[24], transcriptomic profiling[25], and mass spectrometric analysis of mosquito protein expression[26]. However, thus far the level of accuracy, high cost, and/or need of highly trained users suggest that they might not be suitable for application in the field.

In addition to age, identification of mosquito species is crucial for estimation of malaria transmission dynamics. In Africa, the bulk of malaria transmission is carried out by members of the *Anopheles gambiae* sensu latu and *Anopheles funestus* sensu latu species compleses[27]. The *An. gambiae* s.l. complex includes several morphological identical sibling species that can only be distinguished by molecular analysis[28–30]. Despite being morphologically identical, members of this group vary significantly in behaviour, transmission potential, and response to vector control measurements[31]. For example, two major vectors in the *An. gambiae* s.l. group, *An. arabiensis* and *An. gambiae*, can differ in their propensity to enter and rest in houses, their host species choice, breeding conditions, resistance to insecticides, and tolerance to dry climates[6,32,33]. Currently, *An. gambiae* s.l. species are best distinguished by polymerase chain reaction (PCR) methods[30,34–36], which are time-consuming and still relatively costly, and can thus only be carried out on a subsample of mosquitoes collected during typical entomological surveillance conducted by many agencies in Africa. Alternative techniques have been developed such as isoenzyme electrophoresis[37] or chromatography of cuticular components[24], but these are also very laborious and have weak discriminatory power[38].

Non-PCR-based methods often rely on structural and chemical differences in the cuticle to discriminate insects according to their species and other traits. In particular, near-infrared spectroscopy (NIRS) has been evaluated as a general strategy for examining insects since it does not require reagents and holds promise as a fast, practical, non-destructive, and cost-effective method for entomological surveillance. The results obtained to date have proved that the chemical composition of mosquitoes and other insects not only changes between species[39–41], also across different age[40,42–45], according to resistance to insecticides[46] and in the presence of an infectious

agent[47,48]. While promising, the NIRS typical approach has certain drawbacks. As it employs the most energetic portion of the infrared spectrum, the absorption bands are generated by two indirect processes: overtones (a vibration excited at a multiple of the fundamental frequency) and combinations (two or more fundamental vibrations excited simultaneously). Both processes are more incoherent and less frequent than the absorption of light by fundamental vibrations, so their absorption bands are wide and weak. As a result, NIR spectrum of a mosquito, formed by a combination of dozens of these bands, consists of a few features standing out against a background of continuous absorption[49]. Also, most NIRS analyses use a dispersive method to collect the absorption spectra from insects, so the reflectivity of the sample is not controlled and the intensity of the bands of the spectrum depends on how the mosquito is placed in the spectrometer. In addition, the results are normally analysed using Partial Least Squares (PLS) regression, which is prone to over-fitting (i.e. the production of a model that corresponds too closely to a particular set of data and may therefore fail to predict future observations reliably)[50]. This problem commonly arises when the number of samples is relatively small, and the number of variables is large.

Here we tested if these limitations can be overcome by shifting the measurement range (25,000–4,000 cm^{-1}) to the mid-infrared region (4,000–400 cm^{-1}), employing an attenuated total reflectance (ATR) device to assess the mosquitoes, and modelling the results with supervised machine learning. The mid-infrared absorption spectrum of a mosquito contains a set of discrete well-delineated bands that depend on the fundamental vibrations of the molecules present in the cuticle, providing a wealth of information not present in the near-infrared range, where it is not possible to capture the contributions of different biochemical components of the mosquito to the spectrum and their variations among mosquitoes with different attributes, as shown in *Aedes aegypti* and the diptera *Culicoides sonorensis*[51,52]. However, since the mid-infrared spectral bands are affected in non-trivial ways by the development of a mosquito and the changing composition of the cuticle, it is not possible to predict traits by simply monitoring changes in band intensities[51].

Here, we show that the use of supervised machine learning[53] allows the determination of the age and species of two major malaria vectors, *An. arabiensis* and *An. gambiae*, from the information contained in their mid-infrared spectra. This is possible because machine learning, unlike standard statistical approaches, can recognise the complex relationships in these traits (mosquito species and mosquito age) and disentangle them from other irrelevant variation[54–56]. Using this approach, we are able to reconstruct simulated age distributions of mosquito populations with unprecedented reliability. The technique we propose here is time efficient (an analysis takes less than one minute per mosquito), economical, and requires neither reagents nor highly trained operators. It also represents a novel approach to the analysis of insects using spectroscopic techniques, solving some previous drawbacks, and accelerating progress towards the establishment of infrared spectroscopy as a routine approach for mosquito surveillance and evaluation of interventions.

Methods

Mosquito rearing, blood feeding, and processing

Anopheles gambiae s.s (Kisumu strain) and *An. arabiensis* (Ifakara strain) mosquitoes were reared under standard insectary conditions of 27 ± 1°C, 70% humidity and a 12-hr light: 12-hr dark cycle at the University of Glasgow. *Anopheles gambiae* s.s (Kisumu strain) mosquitoes were provided by Hilary Ranson (Liverpool School of Tropical Medicine). The *An. arabiensis* (Ifakara strain) colony was initially established in 2008 at the Ifakara Health Institute with individuals from Sagamaganga village (Kilombero District, Morogoro Region, Tanzania)[57], and after a few generations reared at the University of Glasgow. Larvae were fed *ad libitum* on fish pellets (Tetra Pond Pellets, Tetra GmbH, Herrenteich 78, D49324). Pupae were collected from the larval trays and moved into a cage for emergence. Mosquitoes were considered to be in the age category of "Day 0" on their day of emergence from pupa to adult. Upon emergence, adults were fed *ad libitum* on a 5% glucose solution supplemented with 0.05% (w/v) 4-aminobenzoic acid (PABA).

In order to produce mosquitoes with the same age and different physiological conditions, cages with mosquitoes of the same age (where pupae were added on the same day) were blood fed human blood and membrane feeders at different days after emergence. An oviposition cup was then introduced 2 days after a blood meal to allow egg laying. Mosquitoes under three types of physiological conditions were collected, specifically: mosquitoes that had just received a blood meal (blood fed), mosquitoes that developed eggs as they received a blood meal two days before collection (gravid) and mosquitoes that laid eggs as they received a blood meal four days before collection and had the chance to lay eggs on an oviposition cup for two consecutive nights (sugar fed). Blood feeding was provided to each cage every 6 days. Thus, mosquitoes living 6 or more days after their first blood meal underwent multiple gonotrophic cycles.

Human blood was obtained from the Glasgow and West of Scotland Blood Transfusion Service. Ethical approval for the supply and use of human blood was obtained from Scottish National Blood Transfusion Service committee for governance of blood and tissue samples for non-therapeutic use, and Donor Research (submission Reference No 18~15). Whole blood from donors of any blood group was provided in Citrate-Phosphate-Dextrose-Adenine (CPD-A) anticoagulant/preservative. Fresh blood was obtained on a weekly basis.

Upon collection, mosquitoes were transferred into a cup and killed with a cotton soaked with chloroform placed on top of the cup for 30 minutes. Dead mosquitoes were then transferred into a tube over a layer of cotton and silica gel desiccant. The vial was then immediately stored at 4°C. Since it takes one day to dry in silica *Anopheles gambiae* mosquitoes and two days

An. arabiensis, both species were stored prior to measurement for at least three days.

Spectral data acquisition

Dried specimens were laid on their sides on the ATR diamond so that the surface of the diamond was mainly covered by the insect's head and thorax to avoid as far as possible measuring the contents of the abdomen (Figure 1). The wings and limbs were not removed and were used to help position the mosquito. Pressure was then applied by the anvil of the ATR and the spectrum was measured using a dry-air purged Bruker Vertex 70 spectrometer (Bruker Corporation, Billerica, Massachusetts, USA) equipped with a Globar lamp, a Deuterated Lanthanum α Alanine doped Tri-Glicine Sulphate (DLaTGS) detector, a potassium bromide (KBr) beamsplitter, and a diamond ATR accessory (Bruker Platinum ATR Unit A225). Final, noiseless spectra were produced after averaging 16 scans taken at room temperature between 400 and 4,000 cm-1 with 1 cm^{-1} resolution. Mosquito spectra with low intensity or a significative atmospheric intrusion (Figure 2) were discarded automatically using *Loco Mosquito* 5.0, a custom program written in Python 3.6 (see Software availability section). This program discarded unsuitable spectra by measuring the average absorbance of the plateau in the mosquito spectra between 400 and 500 cm^{-1} and the smoothness of the region between 3,500 and 3,900 cm^{-1} (to detect water and CO_2 spectra intrusion).

Machine-learning analysis

A supervised machine-learning approach was used to map the pre-selected 17 wavenumbers (see Spectroscopic Method subsection in Results) to mosquito species (either *An. gambiae* or *An. arabiensis*) and to mosquito age. In both cases, a classification approach was used. The age classes selected were mosquito ages 1, 3, 5, 7, 9, 11, and 15 days, which allowed acceptable per-age accuracy while improving on current binary cut-off of 4 days based on oviposition (and assuming no pre-gravid behaviour). These age classes were chosen as a compromise between granularity of the predictions and model performance.

Mosquito species and ages were treated in separate models to increase accuracy. To identify the algorithms most suited to the identification of either mosquito species and age class, we first compared the baseline performance of k nearest neighbours (kNN), logistic regression (LR), support vector machines (SVM), random forests (RF), and gradient boosted trees (XGB) using 5-fold cross-validation (Figure 3). This range of parametric (LR, SVM) and non-parametric (kNN, RF, XGB) models offer different data representation schemes using Euclidian distance (kNN), linear relationships (LR, SVM), and ensemble decision trees (RF, XGB). For species and age class identification, XGB and LR, respectively, were then selected for further optimization. The full dataset—comprising 2,536 mosquito spectral features (details in Table 1) and their corresponding species or age labels—was sampled at random to generate a hold-out validation set stratified according to predicted age classes for each species (see below). The remaining samples were then repeatedly (10 rounds) split in random stratified training and test sets (10 folds). Model optimization involved a further 70%/30% random stratified splitting scheme on each of the training folds, and algorithms were trained with a broad range of parameter combinations, and the best settings for each train set retained.

Each optimised model's accuracy was then calculated against the corresponding test set. The 100 resulting trained models were then ranked according to their accuracy scores, and the best 10 retained and predictions bagged for evaluation of their predicted labels (age or species) against the true labels. All machine learning was performed in Python 3.6 using scikit-learn 0.19, XGBoost 0.82, and corresponding plotting using seaborn 0.9.

Age-structure modelling

To illustrate the utility of our approach for field-based surveys of *Anopheles* populations, and to assess whether they could be used to measure the impact of vector control interventions in the field, we simulated age structures of *An. gambiae* and *An. arabiensis* using a simple age structure population model. Here, age corresponds to days. Specifically, the number of mosquitoes N surviving to from age t to t+1 was modelled as a binomial function: $N_{t+1} \sim \text{binomial}(N_t, s)$; where N_t is the total number of mosquitoes alive at age t+1 and s is the probability of daily survival. The daily survival rate was based on literature values, i.e., for *An. gambiae* $s = 0.91$[58] and for *An. arabiensis* $s = 0.82$[16]. For the age structure of the populations under a theoretical intervention regime, we assume that the intervention quadruples the mortality rate of both species from day 3 onwards. This emulates a scenario where mosquitoes encounter an insecticide-treated bednet for the first time at day 3, when they start feeding.

Each age class was generated by sampling the full dataset in the proportions calculated from the above simulated age-structured populations. A continuous probability distribution was then

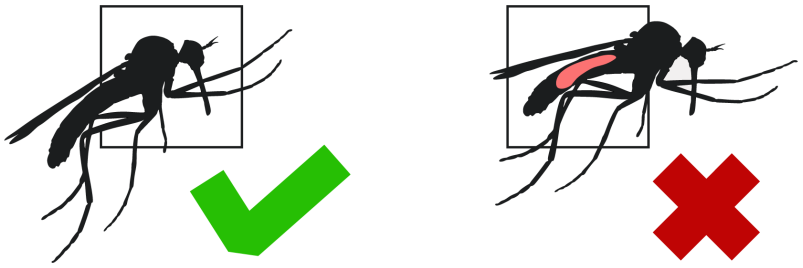

Figure 1. Best position of the mosquito on the ATR crystal. The correct way to place a mosquito on the ATR crystal (left) is to cover the surface with the head and chest. The wrong way (right) is by centring the abdomen on the crystal.

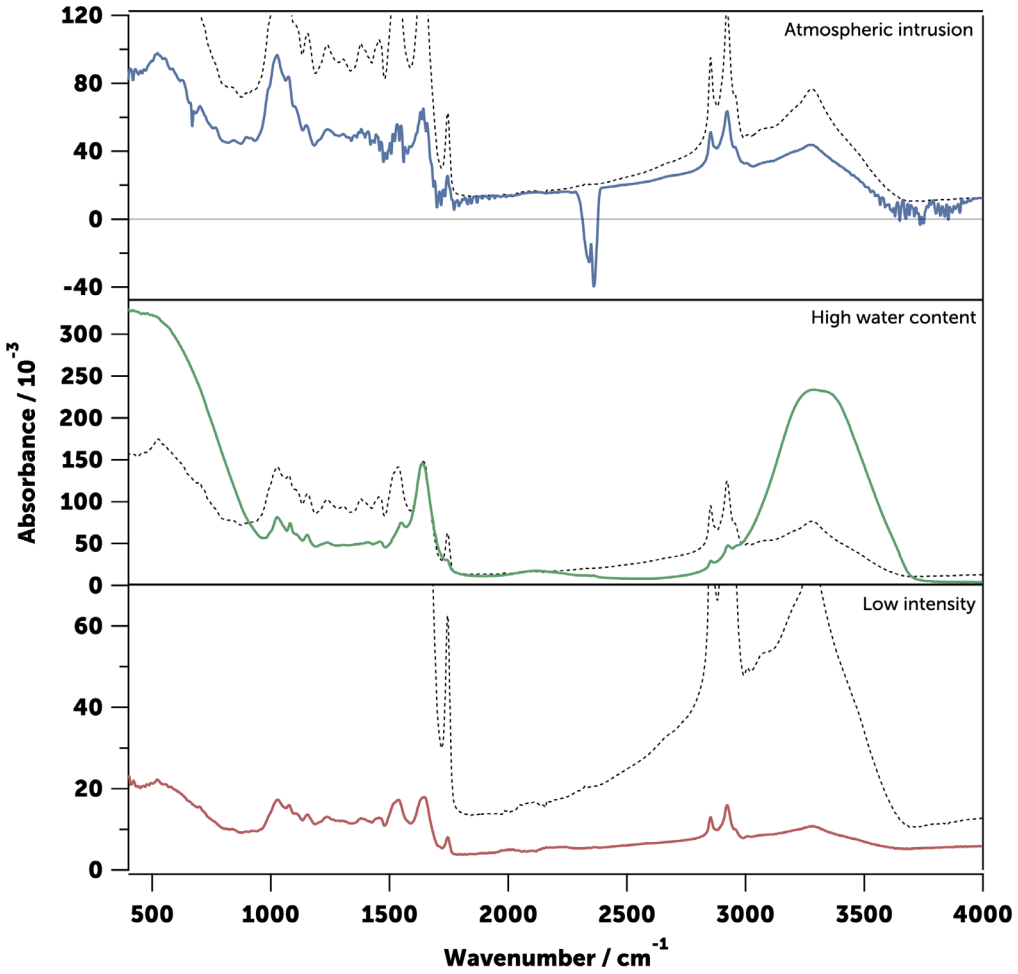

Figure 2. Common experimental errors during the measurement of the infrared spectrum of a mosquito using ATR-FTIR spectroscopy. Above, blue: Spectrum with a significant atmospheric intrusion. Centre, green: *An. gambiae* mosquito with high water content. Below, red. Spectrum with poorly defined features due to low intensity, caused by the displacement of the mosquito during the measurement. All spectra are compared to a correct spectrum of a mosquito shown with a black dashed line.

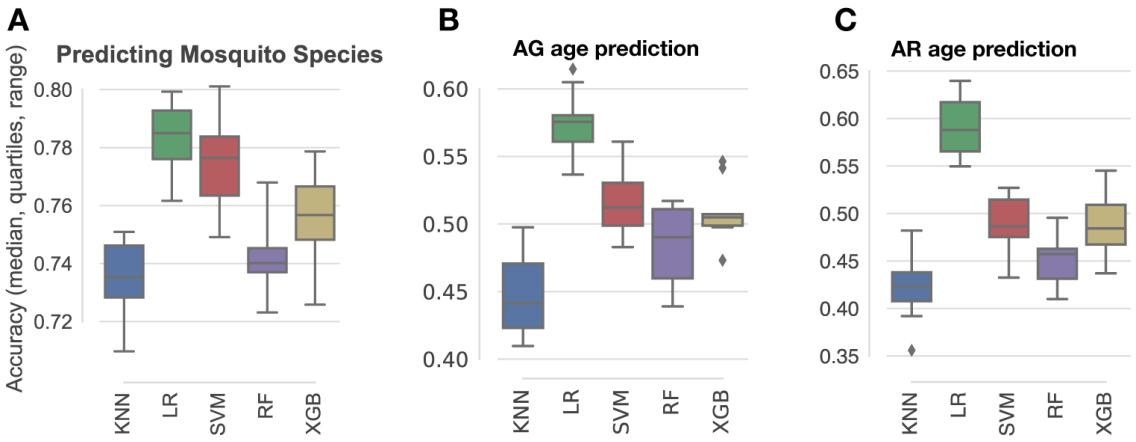

Figure 3. Comparison of the baseline pre-optimisation performance of 5 supervised machine learning algorithms for the prediction of mosquito species (**A**), *An. gambiae* age (**B**), and *An. arabiensis* age (**C**). Each classifier was run with 5-fold cross-validation on a training subset sampled random representing stratified 70% of the full dataset and tested against the remaining 30%. No model parameter optimization was performed at this stage (please see selected models post-optimisation in Figure 13, Figure 14, and Figure 16). KNN, k Nearest Neighbours; LR, logistic regression; SVM, support vector machines; RF, random forests; XGB, gradient boosted trees with XGBoost.

Table 1. Number of mosquitoes of each species and status that have been measured.

Anopheles arabiensis											
											Totals
Age/days	**1**	**3**	**5**	**7**	**9**	**11**	**12**	**13**	**15**	**17**	
Gravid	0	57	61	41	66	70	52	90	33	80	550
Sugar-fed	42	43	65	67	84	67	0	39	41	16	464
Totals	42	100	126	108	150	137	52	129	74	96	1014

Anopheles gambiae																		
																		Totals
Age/days	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	
Gravid	0	0	0	47	45	51	43	40	37	39	35	89	34	0	45	61	39	605
Sugar-fed	160	63	65	62	54	52	59	53	53	44	41	32	27	44	44	24	40	917
Totals	160	63	65	109	99	103	102	93	90	83	76	121	61	44	99	85	79	1522

fitted to the true and predicted discrete age distributions to better generalize our discrete model predictions to an exponentially decreasing age structure using a half-logistic probability function as

The half-logistic distribution is well-suited for fitting survival data[59,60]. Age distributions were compared using the Kolmogorov-Smirnov statistic on 2 samples, a two-sided test for the null hypothesis that 2 independent samples are drawn from the same continuous distribution.

Estimation of the light penetration distance in a mosquito

The depth of light penetration for ATR measurements de-pends on the wavelength λ and angle of incidence of light θ, and on the refractive indices of the mosquito, n_2, and the ATR crystal, n_1:

$$d = \frac{\lambda n_1}{2\pi\sqrt{\sin^2\theta - (n_2 - n_1)^2}}$$

Taking into account that, according to the specifications of the ATR accessory[61], the incidence angle is θ = 45º, the Sellmeier equation[62] for diamond[63] (λ in μm):

$$n^2 - 1 = \frac{4.3356\lambda^2}{\lambda^2 - 0.1060^2} + \frac{0.3306\lambda^2}{\lambda^2 - 0.1750^2}$$

and a Cauchy equation $n(\lambda) = A + B/\lambda^2$, with A = 1.517 and B = $8.80 \cdot 10^{-3}$ μm^2 for insect chitin[64]. The results for the MIR region are shown in Figure 4.

Results

Mosquitoes preparation

A 'field-friendly' protocol to kill and store mosquitoes for infrared (IR) spectroscopy was established. In brief, laboratory-reared female *An. gambiae* and *An. arabiensis* mosquitoes of different ages and physiological states were killed by exposure to chloroform for 30 minutes. As chloroform evaporates and does not interact with the mosquito cuticle, the IR spectra were not affected by this chemical (Figure 5). This method, also used before[40,43,45], is more practical in the field than killing mosquitoes with CO_2 or by freezing them at -20°C. Dead mosquitoes were then stored in 20 ml transport tubes with silica gel to dry them out[65]. Removal of water from samples is essential, as it uncovers parts of the IR spectrum that would otherwise be hidden by the intense IR absorption of water (Figure 6). Water IR absorption bands disappeared from *An. gambiae* and *An. arabiensis* mosquitoes after storage with silica gel at 4°C for one and two days, respectively (longer in a *An. arabiensis* due to its higher body water content)[66]. In addition, this drying method preserved mosquitoes from decomposition for more than 10 days (Figure 7). Alternative drying methods such as desiccating specimens in an oven at 80°C were shown to affect IR spectra, disrupting specially the peaks associated with lipids (Figure 6), and therefore not used.

Spectroscopic method

The far- (30–400 cm^{-1}), mid- (400–4,000 cm^{-1}), and near-infrared (4,000–10,000) regions of mosquito spectra were compared (Figure 8). The far- and near-infrared regions were essentially featureless in dried mosquitoes, unlike the NIR spectra previously published[40,43,45,47,67] which show the intense signals of liquid water when specimens were not dried (Figure 9). However, the mid-infrared region showed a large number of well-defined intense peaks, which are easily identifiable as coming from the chemical components of the cuticle (Table 2). Three different IR spectral sampling techniques were investigated: diffuse reflectance, transmission, and attenuated total internal reflection (ATR, see Spectral data acquisition in Methods). ATR spectroscopy produced the best-defined and most reproducible spectra in the mid-IR region (Figure 10). ATR also allowed the measurement of different parts of the mosquito body (*e.g.*, head or abdomen) that have slightly different IR spectra (Figure 11). It also had superior signal-to-noise ratios allowing acquisition of the spectra in 45 seconds. Raw spectra data is available as Underlying data[68].

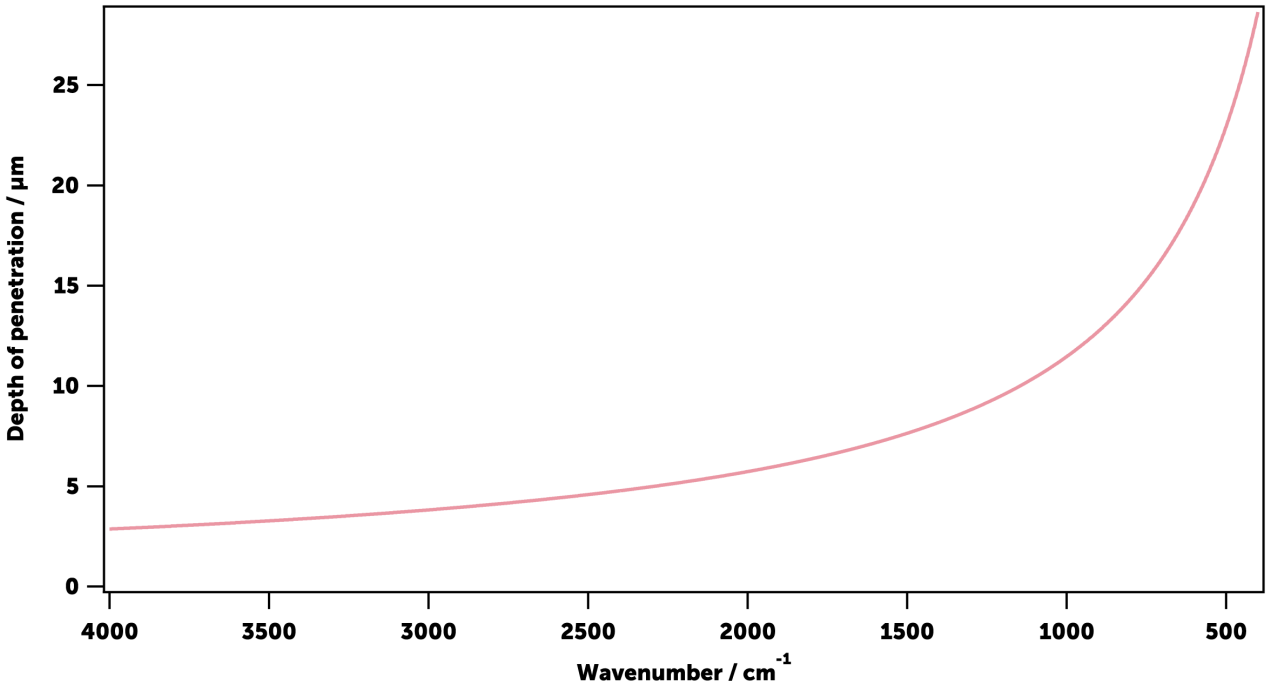

Figure 4. Estimated depth of penetration of the ATR evanescent wave in the mosquito sample.

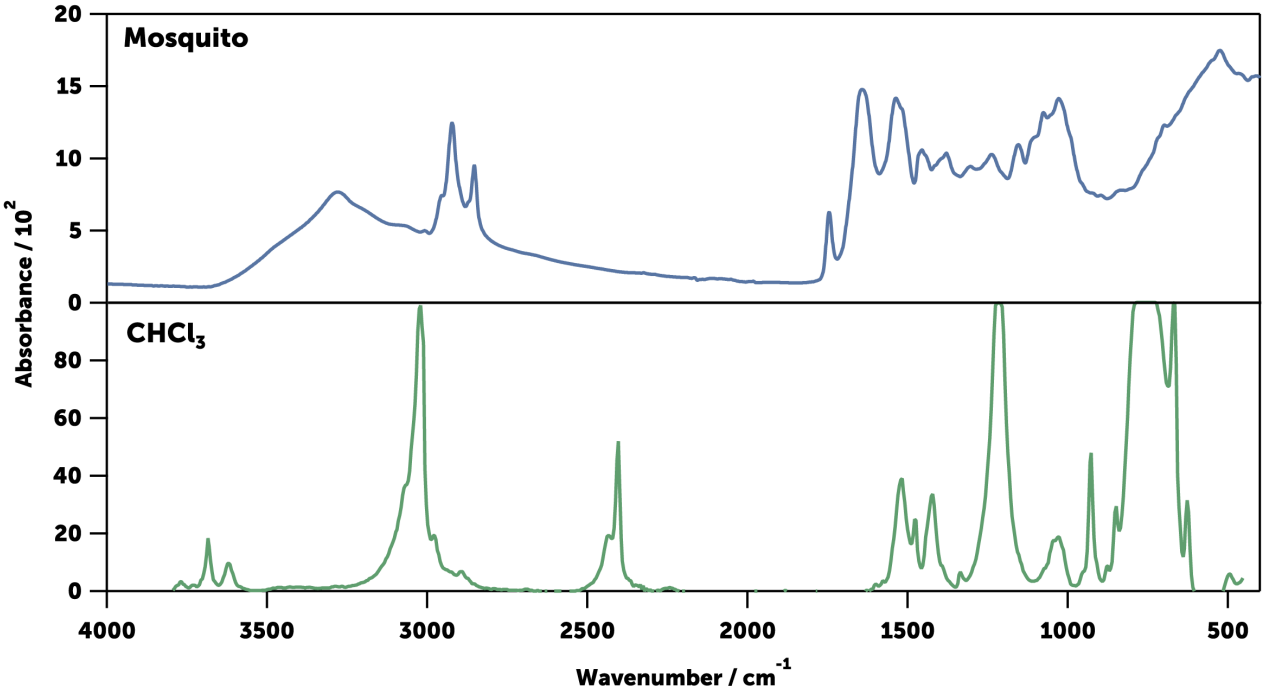

Figure 5. Mid-infrared absorption spectra of a typical mosquito (*An. gambiae*, gravid, 9 days old, top) and liquid chloroform (bottom). Note the absence of the signal of the chloroform employed to kill the mosquito in the insect spectrum, since chloroform rapidly evaporates from the sample and leave no MIR-detectable signals.

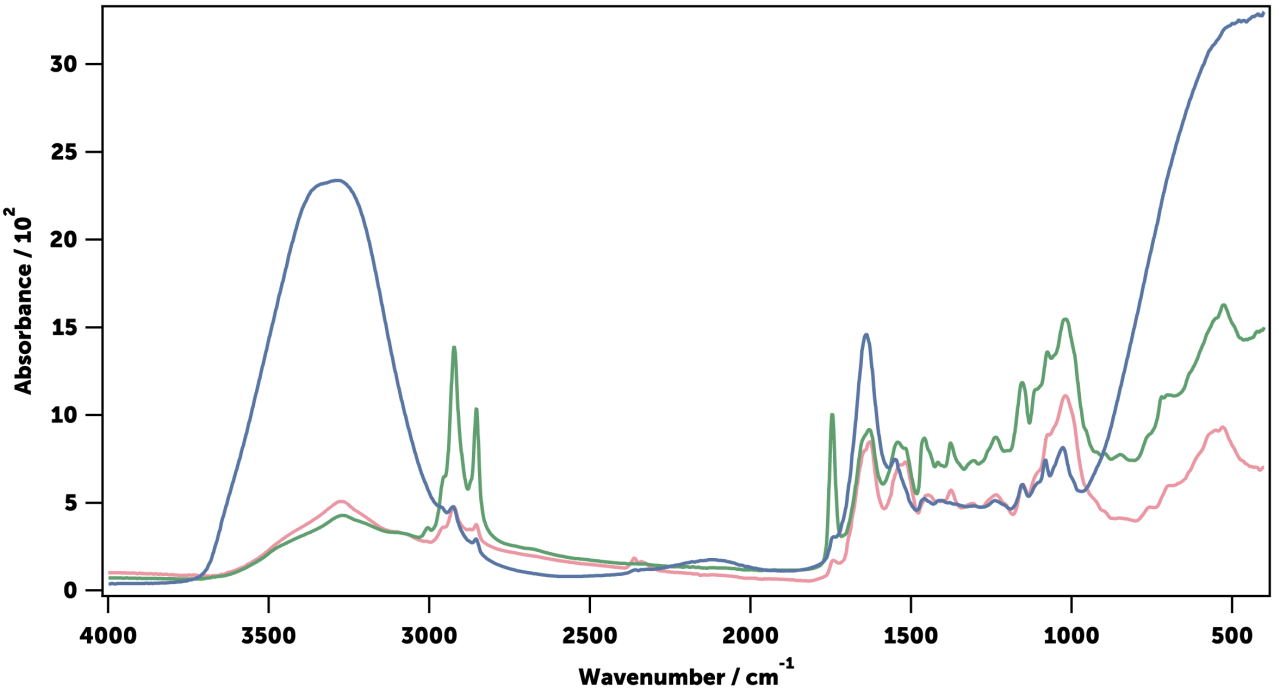

Figure 6. Mid-infrared absorption spectra of a recently killed mosquito (blue), a mosquito dried in a vial with silica (green) and in an oven at 80°C for 60 minutes (pink). All mosquitoes were *An. gambiae*, sugar-fed and 11 days old. A clear loss of detail can be observed in the oven dried sample due to heating.

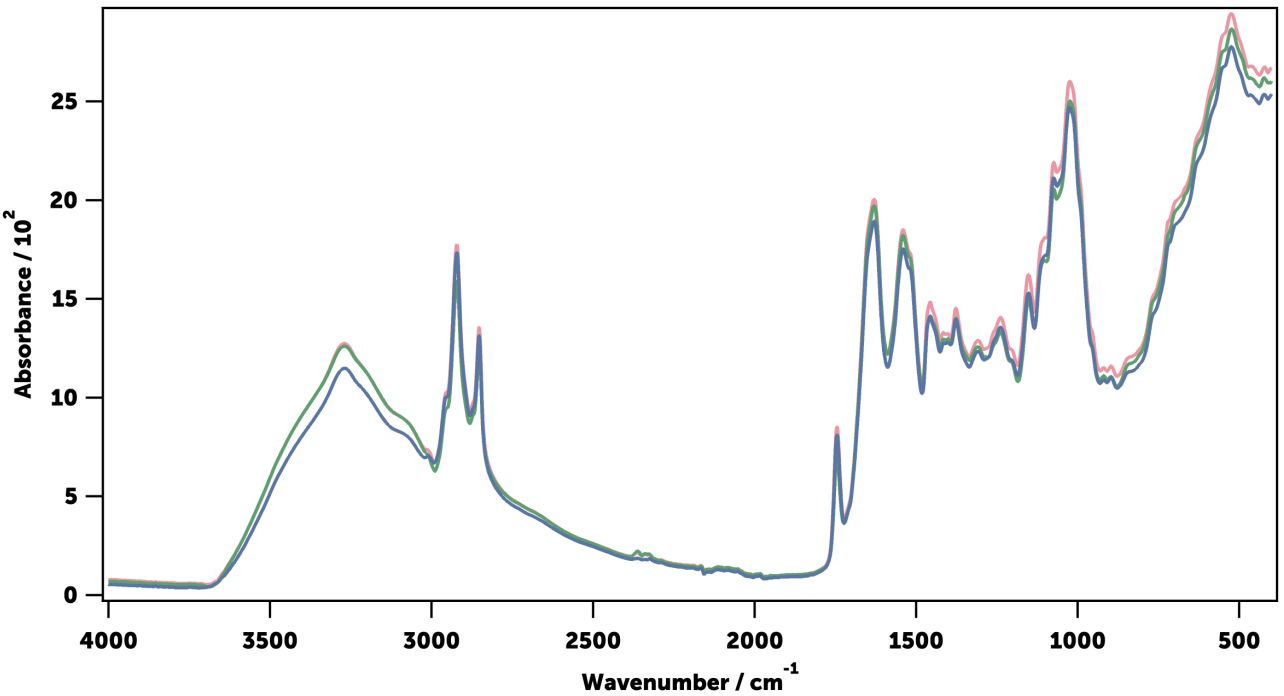

Figure 7. Effect of the storage time on the averaged mid-infrared spectra of 30 sugar-fed 17-day-old An. gambiae mosquitoes. 3 days (blue), 6 days (green), and 11 days (pink) after collection.

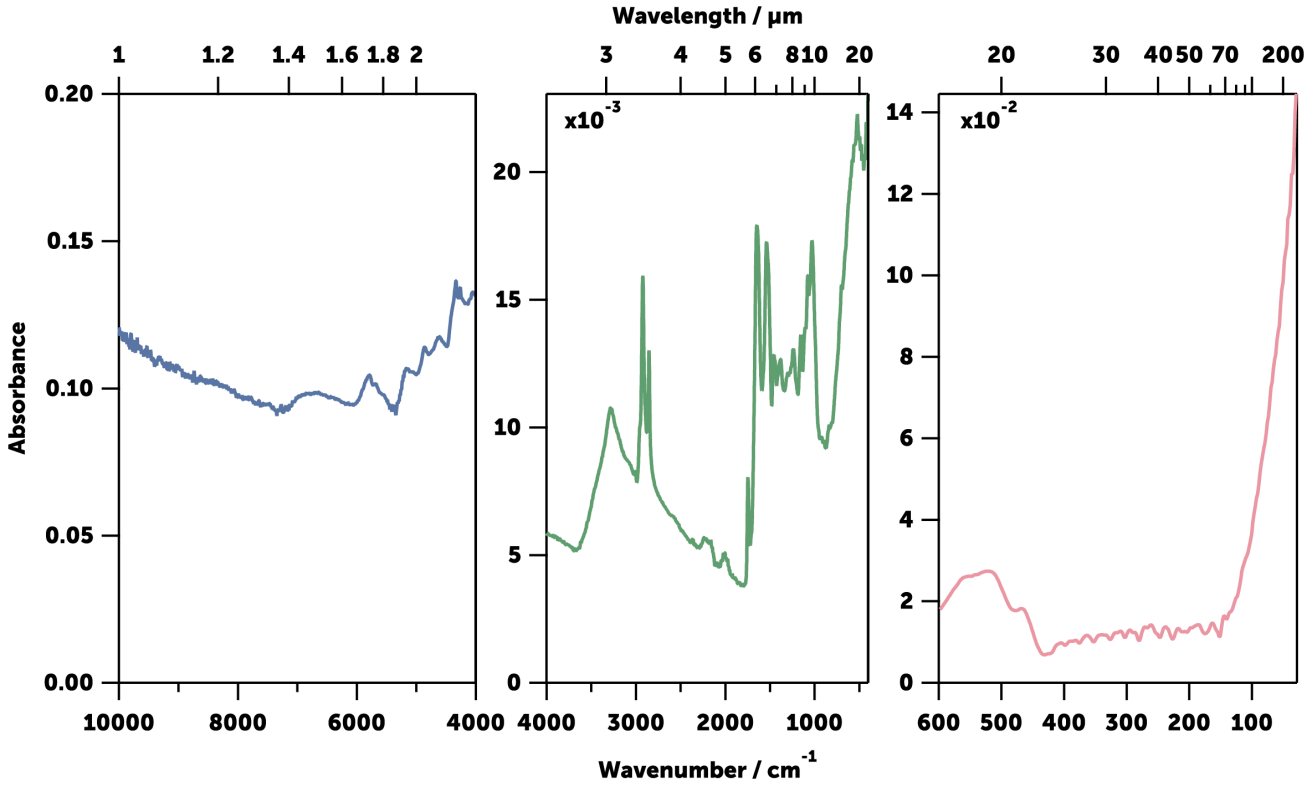

Figure 8. Typical near- (left, blue), mid- (centre, green), and far-infrared (right, pink) spectra of an *An. gambiae* mosquito. The near-infrared spectrum was collected using diffuse reflectance infrared spectroscopy, while the mid- and far-infrared spectra were obtained using ATR.

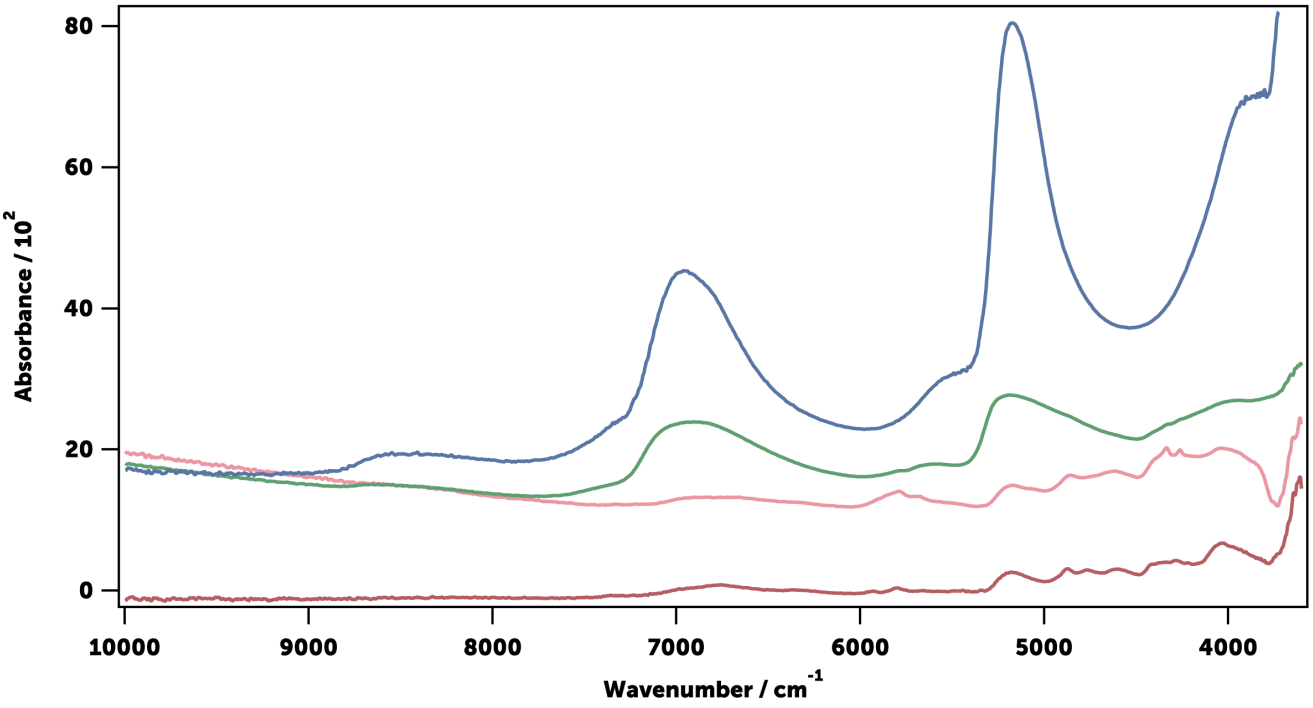

Figure 9. Near-infrared diffuse reflectance spectra of water (blue), an undried *An. gambiae* mosquito (green), a dried mosquito (pink), and chitin (red).

Table 2. Assignment of the selected wavenumbers shown in Figure 12[51,69].

Wavenumber	Bond	Reference
3856	*	
3400	O-H	Mosquito moisture
3276	N-H	Chitin, proteins
2923	$C-H_2$	Proteins, waxes
2859	$C-H_2$	Proteins, waxes
1901	*	
1746	C=O	Proteins, waxes
1636	C=O	Proteins, chitin
1539	O=C-N	Proteins, chitin
1457	$C-CH_3$	Wax, proteins
1307	C-N	Proteins, chitin
1154	C-O-C	Chitin, waxes
1076	C-O	Chitin
1027	C-O	Chitin
880	*	
526	C-C	Proteins, chitin
401	*	

* Wavenumbers selected as indicators of overall spectra intensity and offset.

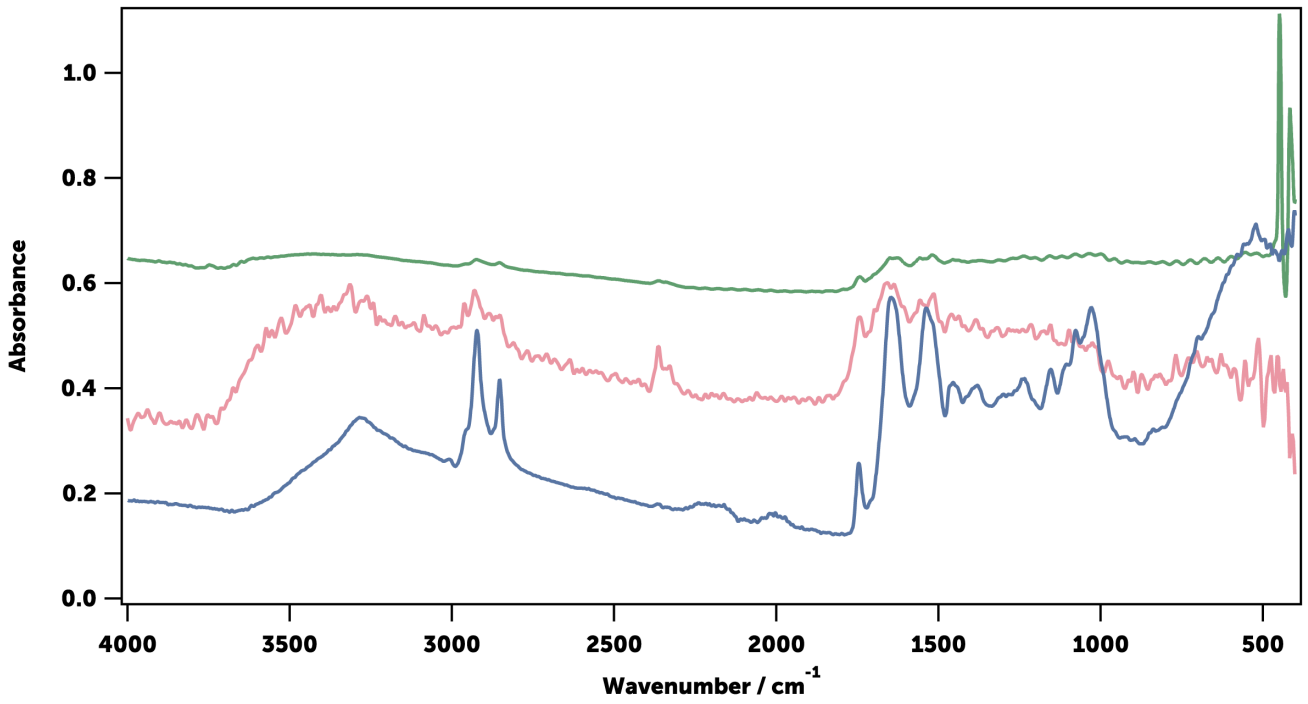

Figure 10. Typical ATR (blue, scaled Abs x32), diffuse reflectance (pink), and transmission (green) mid-infrared spectra of a mosquito. The transmission spectrum was taken using ZnSe windows. Its vertical offset is due to the reflection of a part of the light because of the difficulty in controlling the angle of the cell windows with the mosquito inside. The ATR, diffuse reflectance, and transmission spectra are the result of an average of 16, 120, and 80 scans, respectively.

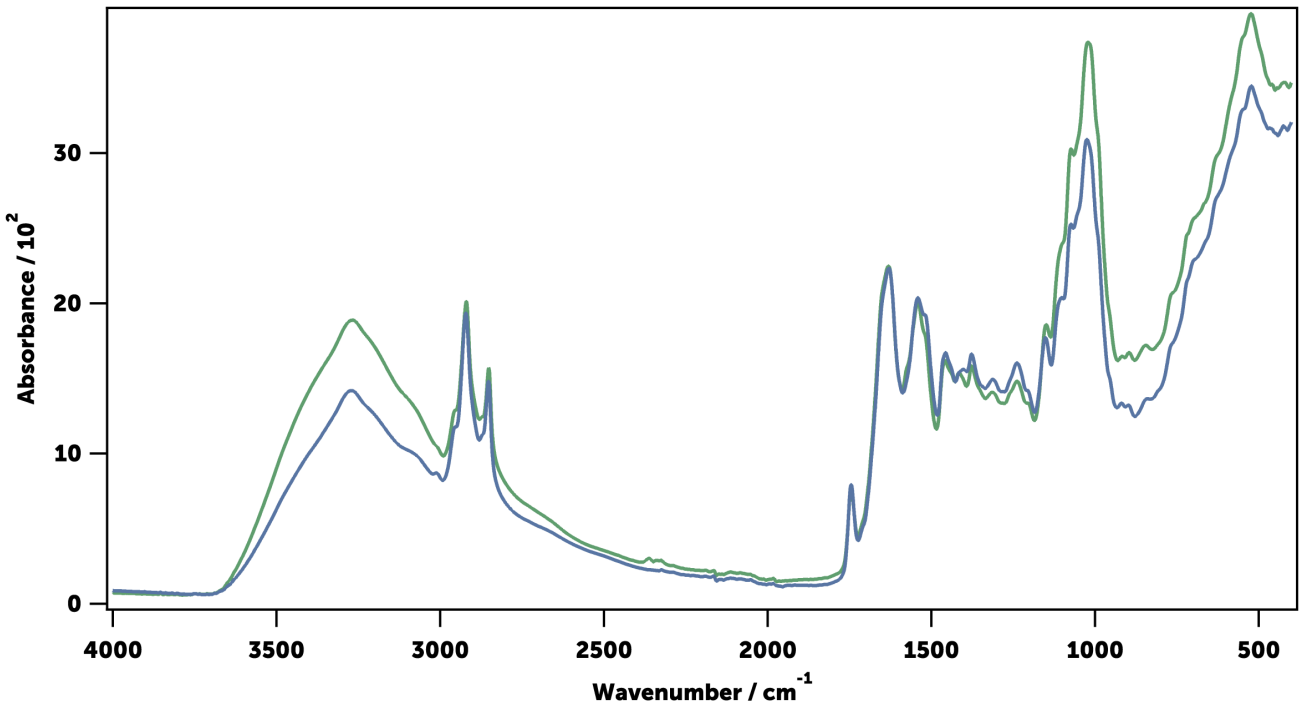

Figure 11. Mid-infrared absorption spectra of the head and thorax (blue) and abdomen (green) of a sugar-fed, 17-day-old, An. gambiae mosquito.

It was estimated that by using the ATR sampling technique in the mid-IR, the light penetrates the sample by 3–10 µm up to about 1000 cm^{-1}, and then up to 22 µm within 1000 and 400 cm^{-1} (see Estimation of the light penetration distance in a mosquito in Methods). As the cuticle of a mosquito is approximately 2–5 µm thick[22,70], the measured spectra encompass the outer shell and part of the interior of insects. As the cuticle is mainly composed of chitin, proteins, and lipids, spectra associated with these substances were individually compared with the whole-mosquito spectra (Figure 12) to allow the assignment of the main vibrational modes of the mosquito cuticular constituents to each element (Table 2). As the cuticular chemical composition is known to change with species and age[71,72], so too are the relative magnitudes of these vibrational bands. To quantify this change, 17 wavenumbers in the MIR spectrum were selected corresponding to 13 well-defined vibrational absorption peaks (contributed in different proportions by the three main constituents) and 4 troughs (that provide information on spectrum intensity and offset). These 17 wavenumbers were then used for training machine learning models (see below).

Mosquito species determination

To develop a MIRS-based method to determine the age and species of *An. gambiae* and *An. arabiensis*, mosquitoes were reared under laboratory conditions (see Mosquito rearing, blood feeding, and processing in Methods) and collected at ages ranging from 1 to 17 days. To model part of the variability typical in the wild, female encompassing a range of physiological states were incorporated in analysis including those that have just taken a blood meal (blood fed), those that had eggs developed in the abdomen (gravid), or that laid eggs but have not blood-fed yet again (sugar fed); mosquitoes undergone either single or multiple gonotrophic cycles depending on their age. In most cases, over 40 mosquitoes per age and physiological condition from each species were analysed (Table 1).

A total of 1,522 *An. gambiae* and 1,014 *An. arabiensis* spectra from different ages and physiological conditions were used to train supervised machine-learning models (see Machine-learning analysis in Methods). Five algorithms were tested on the dataset to predict mosquito species (Figure 3A). This initial approach identified logistic regression (LR) as the most accurate approach. We generated 100 bootstrapped models trained on and tested against different subsets of the data which, when aggregated (bagged), predicted the species identity of *An. gambiae* and *An. arabiensis* with 76.8 and 76.6% accuracy, respectively (Figure 13 A). To increase the accuracy of the prediction while retaining the stability and generalisability afforded by bagging, we selected the 10 best models among them, which achieved 82.6% accuracy (Figure 13 B). These results demonstrate that the MIRS signal is indicative of mosquito species and can be used to distinguish between species in a more time and cost-efficient method, although currently with less accuracy, than standard PCR methods.

Mosquito age determination

After the development of the species-prediction model, a similar supervised machine-learning approach was used to model the chronological age for a given mosquito species. Mosquitoes were screened every second day after emerging as

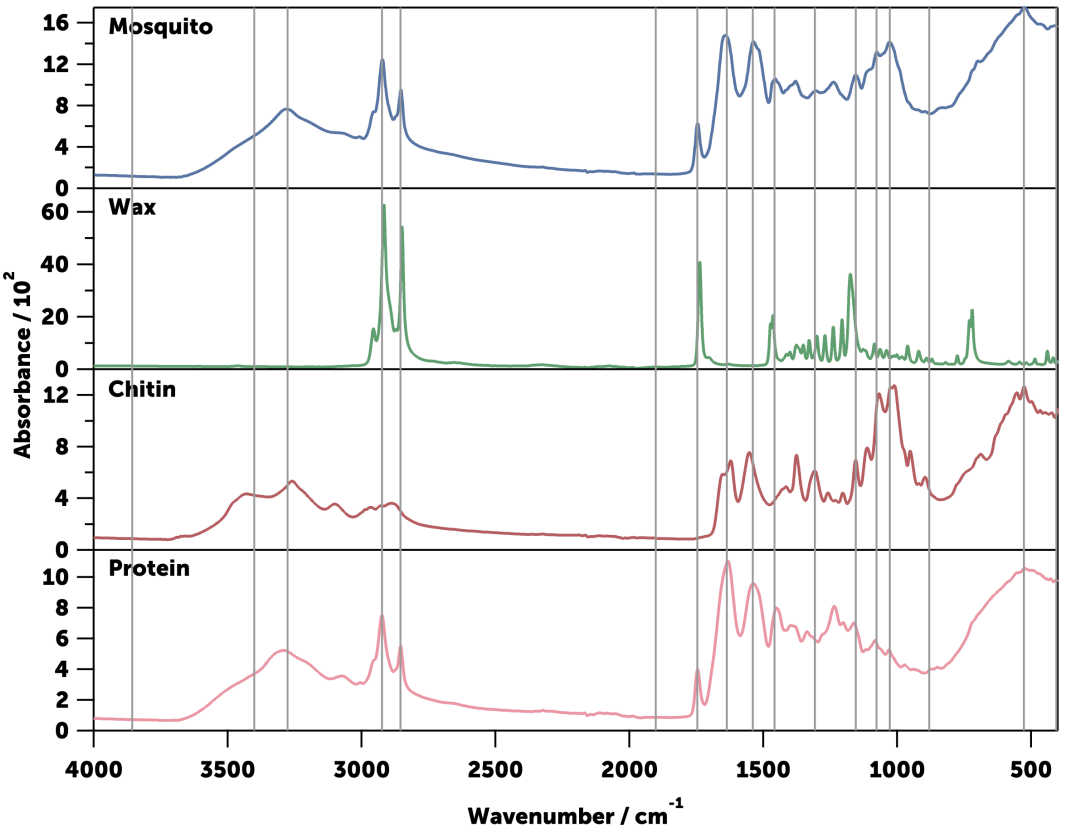

Figure 12. Typical mid-infrared spectrum of an Anopheles mosquito. Shown are *An. gambiae* (gravid, 9 days-old, blue) and its main chemical constituents wax (arachidyl dodecanoate, green), chitin (from shrimp shells, red), and protein (collagen from bovine Achilles tendon, pink). The wavenumbers selected for the machine learning are indicated with a grey line (Table 2).

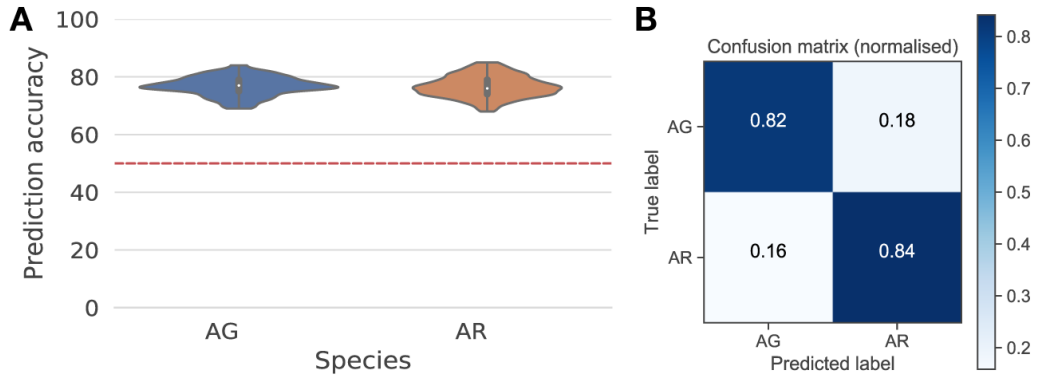

Figure 13. Prediction of mosquito species using mid-infrared spectra. (**A**) Violin plots of the distribution of per species prediction accuracies of 100 models trained on different random stratified subsets (70/30 splits) of the data. The red line shows model prediction accuracy under chance alone (i.e. in the absence of learning). (**B**) Confusion matrix showing the proportion of accurate (diagonal) classification of mosquitoes as either *An. gambiae* (AG) or *An. arabiensis* (AR) using the 10 best logistic regression models (n = 2,536).

adults, and models trained on the same set of 17 wavenumbers as above. The LR model again performed best for both species in correctly mapping wavenumber intensities to mosquito age (Figure 3B and C). To train, optimise, and validate the models, the full dataset was partitioned into an age-structured validation set and retained for later use in population models (see below). The remaining samples were then randomly split into stratified 70%/30% training and test sets for model tuning. The accuracy in predicting each chronological age varied over mosquito lifespan and between species, ranging from an average of 15% to 97% for

An. gambiae and 10% to 100% for *An. arabiensis* (Figure 14). As in previous studies[40,43], it was found that the chronological age of young and old mosquitoes was generally more accurately predicted than intermediate ages, although there were some differences between species. These results suggest that the MIRS-based approach developed here can predict the chronological age of each species from 1 to 15 days old, as well as providing the confidence of prediction for each age class. Furthermore, a trade-off was observed between the granularity of the prediction and its accuracy: models trained on daily scans (not shown) performed worse than if we allowed the mosquitoes to age 2 or 3 days between each scan, suggesting that the ageing of the mosquito cuticle varies between individuals and that the features used for training the models overlap between consecutive age classes (Figure 15).

Predicting mosquito age structure

To monitor the efficacy of vector control interventions in the field, accurately describing the age distribution (*i.e.*, the summary demographic age structure of the local vector species population) is more important than knowing the age of any individual mosquito[46]. Consequently, we tested how well the above age prediction models developed for *An. gambiae* and *An. arabiensis* could reconstruct the known age distribution of mosquito populations. Mosquito populations reflecting anticipated changes in mortality were used under two scenarios: natural mortality and increased mortality due to a theoretical vector control interventions.

Consistent with natural mosquito populations, but unlike our training dataset, field sampling would not produce age-balanced sample sizes, but rather diminishing sample sizes at older age

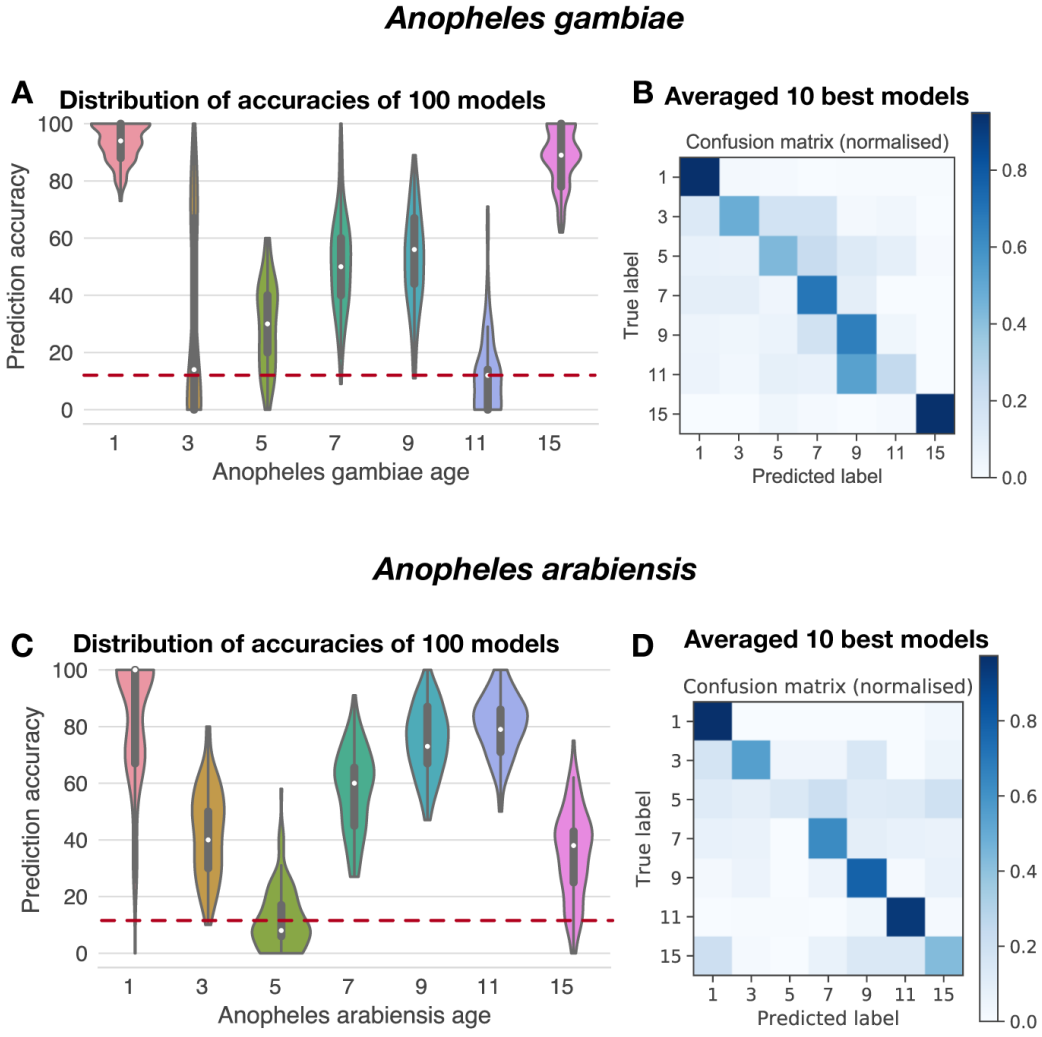

Figure 14. Prediction of *An. gambiae* (**A**–**B**) and *An. arabiensis* (**C**–**D**) age class using mid-infrared spectra. (**A**, **C**) Violin plots of the distribution of per age class prediction accuracies of 100 optimised models. (**B**, **D**) Confusion matrices showing the proportion of accurate (diagonal) classification of mosquitoes as either 1, 3, 5, 7, 9, 11, or 15 days old using the 10 best logistic regression models trained on repeated stratified random subsets using 70% of all mosquitoes sampled, and tested on the remaining 30% (n = 681 for *An. gambiae* and n = 737 for *An. arabiensis*).

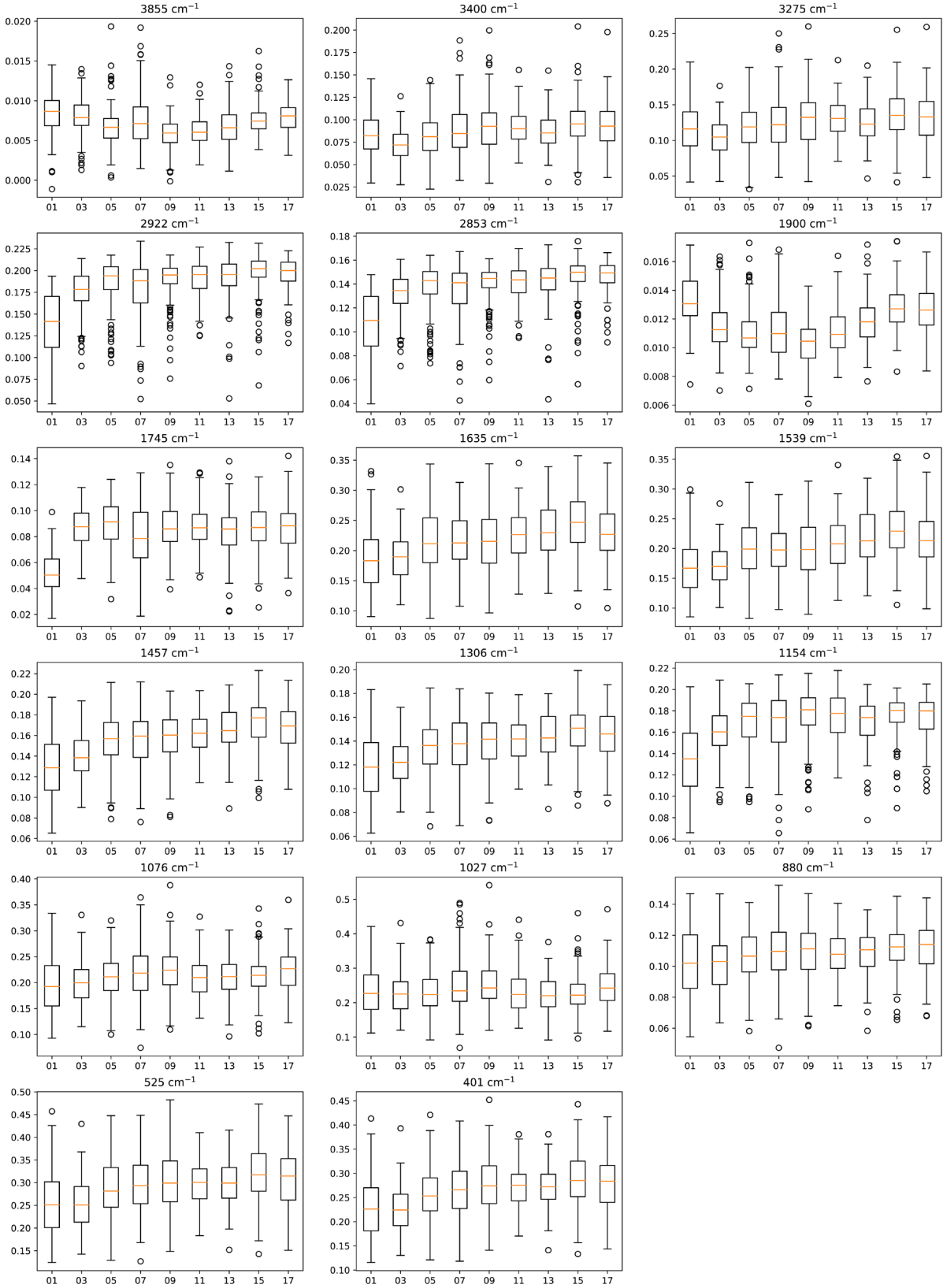

Figure 15. Box-whisker plot containing the measured absorption in each wavenumber and for each age for all the mosquitoes (*An. arabiensis* and *An. gambiae*). The orange lines represent the median absorbance of each age and wavenumber, the limits of the boxes correspond to the interquartile range (IQR) and the whiskers show the lowest datum still within 1.5 IQR of the lower quartile, and the highest datum still within 1.5 IQR of the upper quartile.

classes. Furthermore, it would be highly desirable to use our models to measure the impact of vector control interventions on mosquito-population age structures. However, because no real datasets of a true mosquito population age structure exist, the age structures of *An. gambiae* and *An. arabiensis* were modelled based on their reported average daily mortality[16,58] and assuming an intervention that increased the mortality of adult females four-fold after a first blood meal (~3 days after adult emergence).

In these simulations, a starting population of 1,000 female mosquitoes was used, with the population at each subsequent day being calculated as a proportion of the previous day, with survival rates for each species estimated from reports on field studies (Figure 16A,D)[16,58]. The resulting age-structured populations were then used to randomly sample replicates in the corresponding proportion for each age class used in our MIRS-based prediction models (Figure 15B–F, grey bars; n = 122 for *An. gambiae* and n = 42 for *An. arabiensis*). The models trained above were then used to predict age classes from the MIRS of this age-structured population (Figure 16B–F, orange bars).

To test the ability of those models to reconstruct the age structure of the true population from our predicted age class frequencies, the age structures of the predicted (Figure 16B–F, orange bars) and true sampled populations (Figure 16B–F, grey bars) were modelled with the best fit half-logistic distribution for each species (grey and orange curves in Figures 15 B,C and E,F; see also Age-structure modelling in Methods). The true and predicted age distributions were statistically indistinguishable (Kolmogorov-Smirnov 2-sample test (KS test), $p = 1$ and $p = 0.99$ for *An. gambiae* pre- and post-intervention, respectively; $p = 0.75$ and $p = 0.30$ for *An. arabiensis* pre- and post-intervention, respectively). This approach shows that the algorithm can reconstruct the age structure with good accuracy. Furthermore, our models detected a shift in mosquito age structure consistent with the simulated impacts of the interventions (sampled from true population: KS test $p < 0.0001$ for *An. gambiae* and $p = 0.004$ for *An. arabiensis*; predicted population: $p < 0.0001$ for *An. gambiae* and $p = 0.1$ for *An. arabiensis*), suggesting that this MIRS-based approach holds promise for robust measurement and estimation of the age structure of mosquito vector populations.

Discussion

We developed a straightforward, inexpensive, and rapid method to determine the age and species of large numbers of *An. gambiae* s.l. mosquitoes (*An. gambiae* s.s. and *An. arabiensis*). Based on the supervised machine-learning analysis of their mid-infrared spectra, this method facilitates prediction of mosquito species distribution and survival, two crucial tasks critical to implement and assess malaria control strategies. An advantage of this approach is that in comparison to the current most widely used technique based on dissection, it can determine the whole-age distribution of a mosquito population from the day of emergence until two weeks of age. Although the accuracy of age prediction in the "mid-range" of mosquito life span was not high, by determining the age structure of a population this method could accurately estimate the proportion of mosquitoes within the older and most epidemiologically-important age classes that responsible for malaria transmission.

The use of the mid-infrared spectral region provides some advantages over techniques using near-infrared. Foremost, it is possible to independently quantify the amount of different biochemical components as their vibrational bands appear at different wavenumbers. Furthermore, the MIRS bands are more intense and have much greater definition. In contrast, the near-infrared spectrum of a mosquito is composed of few weak signals (Figure 8) that are typically dominated by the much stronger vibrational overtone and combination bands of water (Figure 9)[46], which is likely more dependent on the mosquito physiological state and environmental conditions than on other mosquito traits, such as species and age.

We have shown that the variation of MIR spectra over mosquito age can be exploited by a machine-learning algorithm to predict the chronological age, and ultimately reconstruct population age structures of two important malaria vector species under simulated conditions of changing mortality risk due to vector control. Our algorithms accurately reconstructed age structures of both *An. arabiensis* and *An. gambiae,* and also detected shifts in mosquito age structure consistent with simulated impacts of interventions. The ability of this proposed technique to predict the age structure of a population suggests that this approach could constitute an efficient tool for monitoring the efficacy of vector control interventions. Future work will include larger datasets used for training in supervised machine learning, comprising field samples with different ecological conditions. The ecological variability of field samples has limited the use of NIRS for age prediction in wild mosquito populations[45]. While the accuracy of MIRS-based approaches may also decline when moving from laboratory-reared to field mosquitoes, we predict that this method will be more robust due to the specific information content and high signal clarity that is obtained in spectra from MIRS. Additional improvements are anticipated by increasing the size and variability of the training set on which mosquito age predictions are validated. This will also facilitate the use of alternative machine learning techniques such as neural networks[73] which may yield even higher accuracy and repeatability.

We have shown that MIRS can discriminate between morphologically identical *An. gambiae* s.l. species with ~83% accuracy. While the observed accuracy of MIRS species prediction is still not comparable to the PCR precision, further work including a larger training set and field samples is expected to increase the overall accuracy of this approach. In addition, the inclusion of other species of the *An. gambiae* s.l. complex will be necessary to implement this technique for field application. However, these laboratory-based results, which included mosquitoes from different ages, physiological conditions, and cohorts, suggest that despite the ecological and life-history traits variation, MIR spectra contain a species-specific signature that the machine-learning algorithm can detect. Indeed, mass-spectrometry studies have shown that different species in the *An. gambiae* s.l. complex have quantitative differences in the cuticular

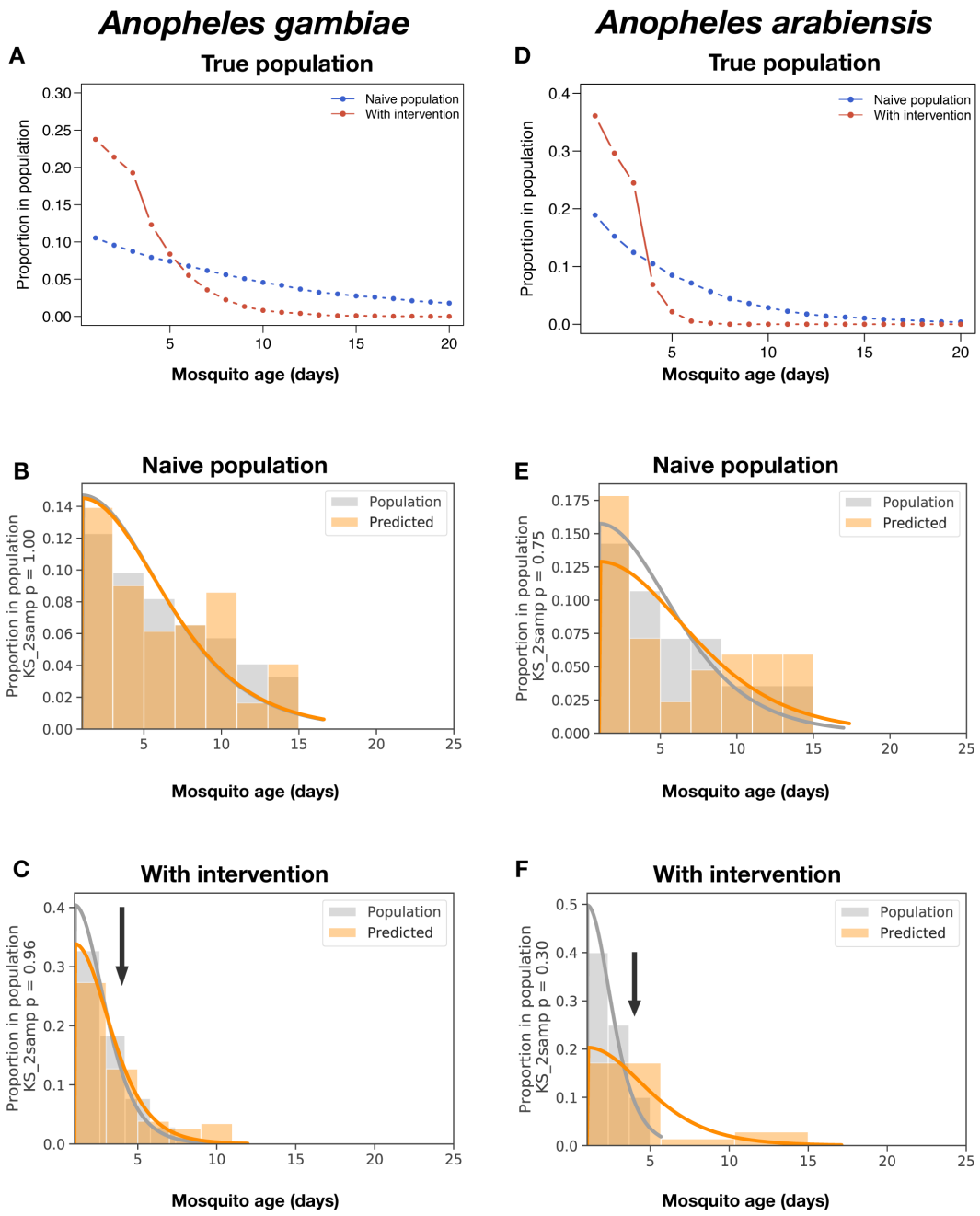

Figure 16. Reconstruction of the age structure of simulated populations of *An. gambiae* and *An. arabiensis* mosquitoes sampled from simulated pre- and post-treatment populations. Population age structures of *An. gambiae* (**A**–**C**) and *An. arabiensis* (**D**–**F**) were generated using an age structure population model assuming survival rates of 0.91 (*An. gambiae*, **A**) or 0.82 (*An. arabiensis*, **D**), under two common scenarios: naive untreated populations (blue lines), and populations in which a simulated vector control program resulted in 4x daily mortality of mosquitoes after day 3 (see Age-structure modelling in Methods for details). The proportions of each age class were extracted from those simulated populations (**A**, **D**), and used to build datasets that are representative of a field-sampled population survey (grey bars in **B**, **C**, **E**, and **F**). The resulting age-structured dataset was then used as the test set for our age-predicting machine learning models (see Figure 3) and compared with the predicted age structure generated from those models (orange bars in **B**, **C**, **E**, and **F**). Finally, we fit a continuous probability distribution to the true (grey curve) and predicted (orange curve) for better generalization of our discrete model predictions to an exponentially decreasing age structure. Population distributions were compared using a 2-sample Kolmogorov-Smirnov test (KS_2samp), reported in the y-axis labels. A - Relative proportion of each age class in a simulated population of *An. gambiae*. B - Estimation of age structure of simulated population from (**A**) using best models from Figure 3B for *An. gambiae* (n = 130). C - Estimation of age structure of simulated population post-intervention from (**A**) using best models from Figure 3B for *An. gambiae* (n = 122). D - Relative proportion of each age class in a simulated population of *An. arabiensis*. E - Estimation of age structure of simulated population from (**A**) using best models from Figure 3D for *An. arabiensis* (n = 42). F - Estimation of age structure of simulated population post-intervention from (**A**) using best models from Figure 3D for *An. arabiensis* (n = 45).

hydrocarbon composition of their cuticle[72], which will affect the MIR spectra.

The biochemical signature obtained by MIRS from the mosquito cuticle provided information on both mosquito species and age. It may therefore be possible to obtain further information on other mosquito traits that alter the cuticular composition. Recently, a new insecticide resistance mechanism has been discovered in *An. gambiae*, which relies on an increased cuticle thickness that in turn reduces insecticide uptake[73]. While this mechanism has been detected by electron microscopy, there are no other methods to measure this new trait, which could have profound epidemiological consequences. In the future, MIRS calibrations including cuticular resistant mosquitoes may be able to identify this insecticide resistant trait. In addition, infection with the *Plasmodium* malaria parasite might be detected by MIRS. Pathogen infection is known to alter mosquito physiology and could directly or indirectly modify their cuticular composition. For example, in the dengue and Zika vector *Aedes aegypti* mosquitoes, an infrared spectroscopy method has recently been developed to detect Zika virus[47], the bacterial endosymbiont *Wolbachia*[52,67], and malaria infection in mosquitoes[41,74].

The accuracy, speed, and generalisability of the MIRS approach presented here shows that this tool holds promise for use in the evaluation of vector control interventions and as triage method when a large number of specimens (>500 -1000) requires to be processed in a rapid fashion. The inclusion of new species, larger sample sizes and field samples with variable ecological conditions is a prerequisite for the application of this technique. It is worth noting that the cost of a portable FTIR MIR spectrometer is ~$20–25,000, which is in the range of quantitative PCR machines used for species determination and/or insecticide resistance monitoring. However, in contrast to PCR analysis, no additional, ongoing costs for reagents and running costs are required once the core equipment is installed. Thus, this approach could be particularly valuable in resource limited settings.

The MIRS method presented here provides rapid and accurate information on *Anopheles* species (82.6%) and reliably characterises mosquito age distribution. However, these results were obtained by training machine learning models with a relatively modest number of mosquitoes (2,536). In future work, it will be possible to generate much larger MIRS datasets and thus train more sophisticated predictive models. Such larger data sets will lend themselves to analysis by more powerful "big data" approaches including deep learning methods that would be expected to improve accuracy considerably beyond this proof-of-principle study. Furthermore, the technique applied to malaria vectors here could also be expanded to other vector-borne diseases such as Zika, dengue, Lyme disease, leishmaniasis, or filariasis. In light of these opportunities, we recommend this method be prioritised for further evaluation.

Data availability

Underlying data

Enlighten: Research Data: Prediction of malaria mosquito species and population age structure using mid-infrared spectroscopy and supervised machine learning. https://doi.org/10.5525/gla.researchdata.688[68]

This project contains the following underlying data:

- DataMosquitoes.zip (zip file containing underlying spectra data)

Software availability

Source code: https://github.com/SimonAB/Gonzalez-Jimenez_MIRS/tree/v1.0

Archived source code: http://doi.org/10.5281/zenodo.2609356[75]

Licence: GNU General Public License v3.0

Acknowledgements

We would like to thank Dorothy Armstrong and Elizabeth Peat for assistance with mosquito rearing and maintenance. We would also like to thank Hilary Ranson for providing the Kisumu colony.

References

1. Bhatt S, Weiss DJ, Cameron E, *et al.*: **The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015.** *Nature.* 2015; **526**(7572): 207–211.
PubMed Abstract | Publisher Full Text | Free Full Text

2. WHO: **World Malaria Report.** 2017.
Reference Source

3. Hemingway J, Ranson H, Magill A, *et al.*: **Averting a malaria disaster: will insecticide resistance derail malaria control?** *Lancet.* 2016; **387**(10029): 1785–1788.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Protopopoff N, Mosha JF, Lukole E, *et al.*: **Effectiveness of a long-lasting piperonyl butoxide-treated insecticidal net and indoor residual spray interventions, separately and together, against malaria transmitted by pyrethroid-resistant mosquitoes: a cluster, randomised controlled, two-by-two factorial design trial.** *Lancet.* 2018; **391**(10130): 1577–1588.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Hawley WA, Phillips-Howard PA, ter Kuile FO, *et al.*: **Community-wide effects of permethrin-treated bed nets on child mortality and malaria morbidity in western Kenya.** *Am J Trop Med Hyg.* 2003; **68**(4 Suppl): 121–127.
PubMed Abstract | Publisher Full Text

6. Pates H, Curtis C: **Mosquito behavior and vector control.** *Annu Rev Entomol.* 2005; **50**: 53–70.
PubMed Abstract | Publisher Full Text

7. Viana M, Hughes A, Matthiopoulos J, *et al.*: **Delayed mortality effects cut the malaria transmission potential of insecticide-resistant mosquitoes.** *Proc Natl Acad Sci U S A.* 2016; **113**(32): 8975–8980.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Beier JC: **Malaria parasite development in mosquitoes.** *Annu Rev Entomol.* 1998;

43: 519–43.
PubMed Abstract | Publisher Full Text

9. Ohm JR, Baldini F, Barreaux P, *et al.*: **Rethinking the extrinsic incubation period of malaria parasites.** *Parasit Vectors.* 2018; **11**(1): 178.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Smith DL, McKenzie FE: **Statics and dynamics of malaria infection in Anopheles mosquitoes.** *Malar J.* 2004; **3**: 13.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Brady OJ, Godfray HC, Tatem AJ, *et al.*: **Vectorial capacity and vector control: reconsidering sensitivity to parameters for malaria elimination.** *Trans R Soc Trop Med Hyg.* 2016; **110**(2): 107–117.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Gillies MT, Wilkes TJ: **A study of the age-composition of populations of *Anopheles gambiae* Giles and *A. funestus* Giles in North-Eastern Tanzania.** *Bull Entomol Res.* 1965; **56**(2): 237–262.
PubMed Abstract | Publisher Full Text

13. Macdonald G: **Epidemiological basis of malaria control.** *Bull World Health Organ.* 1956; **15**(3–5): 613–626.
PubMed Abstract | Free Full Text

14. Macdonald G: **The Epidemiology and Control of Malaria.** Oxford University Press, 1957.
Reference Source

15. Detinova TS: **Age-grouping methods in Diptera of medical importance with special reference to some vectors of malaria.** *Monogr Ser World Health Organ.* 1962; **47**: 13–191.
PubMed Abstract

16. Charlwood JD, Smith T, Billingsley PF, *et al.*: **Survival and infection probabilities of anthropophagic anophelines from an area of high prevalence of *Plasmodium falciparum* in humans.** *Bull Entomol Res.* 1997; **87**(5): 445–453.
Publisher Full Text

17. Polovodova VP: **Age changes in ovaries of Anopheles and methods of determination of age composition in mosquito populations.** *Med Parazitol i Parazit Bolezn.* 1941; **10**(9): 387–395.

18. Yakob L, Yan G: **Modeling the effects of integrating larval habitat source reduction and insecticide treated nets for malaria control.** *PLoS One.* 2009; **4**(9): e6921.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Anagonou R, Agossa F, Azondékon R, *et al.*: **Application of Polovodova's method for the determination of physiological age and relationship between the level of parity and infectivity of *Plasmodium falciparum* in *Anopheles gambiae s.s*, south-eastern Benin.** *Parasit Vectors.* 2015; **8**: 117.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Hoc TQ, Charlwood JD: **Age determination of *Aedes cantans* using the ovarian oil injection technique.** *Med Vet Entomol.* 1990; **4**(2): 227–33.
PubMed Abstract | Publisher Full Text

21. Joy TK, Jeffrey Gutierrez EH, Ernst K, *et al.*: **Aging field collected *Aedes aegypti* to determine their capacity for dengue transmission in the southwestern United States.** *PLoS One.* 2012; **7**(10): e46946.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Schlein Y, Gratz NG: **Determination of the age of some anopheline mosquitos by daily growth layers of skeletal apodemes.** *Bull World Health Organ.* 1973; **49**(4): 371–375.
PubMed Abstract | Free Full Text

23. Gerade BB, Lee SH, Scott TW, *et al.*: **Field validation of *Aedes aegypti* (Diptera: Culicidae) age estimation by analysis of cuticular hydrocarbons.** *J Med Entomol.* 2004; **41**(2): 231–238.
PubMed Abstract | Publisher Full Text

24. Wu D, Lehane MJ: **Pteridine fluorescence for age determination of *Anopheles* mosquitoes.** *Med Vet Entomol.* 1999; **13**(1): 48–52.
PubMed Abstract | Publisher Full Text

25. Cook PE, Hugo LE, Iturbe-Ormaetxe I, *et al.*: **The use of transcriptional profiles to predict adult mosquito age under field conditions.** *Proc Natl Acad Sci U S A.* 2006; **103**(48): 18060–18065.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Sikulu MT, Monkman J, Dave KA, *et al.*: **Mass spectrometry identification of age-associated proteins from the malaria mosquitoes *Anopheles gambiae* s.s. and *Anopheles stephensi*.** *Data Brief.* 2015; **4**: 461–467.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Sinka ME, Bangs MJ, Manguin S, *et al.*: **The dominant *Anopheles* vectors of human malaria in the Asia-Pacific region: occurrence data, distribution maps and bionomic précis.** *Parasit Vectors.* 2011; **4**: 89.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Koekemoer LL, Kamau L, Hunt RH, *et al.*: **A cocktail polymerase chain reaction assay to identify members of the *Anopheles funestus* (Diptera: Culicidae) group.** *Am J Trop Med Hyg.* 2002; **66**(6): 804–811.
PubMed Abstract | Publisher Full Text

29. Cohuet A, Simard F, Toto JC, *et al.*: **Species identification within the *Anopheles funestus* group of malaria vectors in Cameroon and evidence for a new species.** *Am J Trop Med Hyg.* 2003; **69**(2): 200–205.
PubMed Abstract | Publisher Full Text

30. Santolamazza F, Mancini E, Simard F, *et al.*: **Insertion polymorphisms of *SINE200* retrotransposons within speciation islands of *Anopheles gambiae* molecular forms.** *Malar J.* 2008; **7**: 163.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Braack L, Hunt R, Koekemoer LL, *et al.*: **Biting behaviour of African malaria vectors: 1. where do the main vector species bite on the human body?** *Parasit Vectors.* 2015; **8**: 76.
PubMed Abstract | Publisher Full Text | Free Full Text

32. Lyimo IN, Ferguson HM: **Ecological and evolutionary determinants of host species choice in mosquito vectors.** *Trends Parasitol.* 2009; **25**(4): 189–196.
PubMed Abstract | Publisher Full Text

33. Lehmann T, Diabate A: **The molecular forms of *Anopheles gambiae*: a phenotypic perspective.** *Infect Genet Evol.* 2008; **8**(5): 737–746.
PubMed Abstract | Publisher Full Text | Free Full Text

34. Bass C, Williamson MS, Wilding CS, *et al.*: **Identification of the main malaria vectors in the *Anopheles gambiae* species complex using a TaqMan real-time PCR assay.** *Malar J.* 2007; **6**: 155.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Favia G, Lanfrancotti A, Spanos L, *et al.*: **Molecular characterization of ribosomal DNA polymorphisms discriminating among chromosomal forms of *Anopheles gambiae* s.s.** *Insect Mol Biol.* 2001; **10**(1): 19–23.
PubMed Abstract | Publisher Full Text

36. Fanello C, Santolamazza F, Della Torre A: **Simultaneous identification of species and molecular forms of the *Anopheles gambiae* complex by PCR-RFLP.** *Med Vet Entomol.* 2002; **16**(4): 461–464.
PubMed Abstract | Publisher Full Text

37. Cooseman M, Smits A, Roelants P: **Intraspecific isozyme polymorphism of *Anopheles gambiae* in relation to environment, behavior, and malaria transmission in southwestern Burkina Faso.** *Am J Trop Med Hyg.* 1998; **58**(1): 70–74.
PubMed Abstract | Publisher Full Text

38. Al Ahmed AM, Badjah-Hadj-Ahmed AY, Al Othman ZA, *et al.*: **Identification of wild collected mosquito vectors of diseases using gas chromatography-mass spectrometry in Jazan Province, Saudi Arabia.** *J Mass Spectrom.* 2013; **48**(11): 1170–1177.
PubMed Abstract | Publisher Full Text

39. Pickering CL, Hands JR, Fullwood LM, *et al.*: **Rapid discrimination of maggots utilising ATR-FTIR spectroscopy.** *Forensic Sci Int.* 2015; **249**: 189–196.
PubMed Abstract | Publisher Full Text

40. Mayagaya VS, Michel K, Benedict MQ, *et al.*: **Non-destructive determination of age and species of *Anopheles gambiae* s.l. using near-infrared spectroscopy.** *Am J Trop Med Hyg.* 2009; **81**(4): 622–630.
PubMed Abstract | Publisher Full Text

41. Barbosa TM, de Lima LAS, Dos Santos MCD, *et al.*: **A novel use of infra-red spectroscopy (NIRS and ATR-FTIR) coupled with variable selection algorithms for the identification of insect species (Diptera: Sarcophagidae) of medico-legal relevance.** *Acta Trop.* 2018; **185**: 1–12.
PubMed Abstract | Publisher Full Text

42. Perez-Mendoza J, Dowell FE, Broce AB, *et al.*: **Chronological age-grading of house flies by using near-infrared spectroscopy.** *J Med Entomol.* 2002; **39**(3): 499–508.
PubMed Abstract | Publisher Full Text

43. Sikulu M, Killeen GF, Hugo LE, *et al.*: **Near-infrared spectroscopy as a complementary age grading and species identification tool for African malaria vectors.** *Parasit Vectors.* 2010; **3**: 49.
PubMed Abstract | Publisher Full Text | Free Full Text

44. Ntamatungiro AJ, Mayagaya VS, Rieben S, *et al.*: **The influence of physiological status on age prediction of *Anopheles arabiensis* using near infra-red spectroscopy.** *Parasit Vectors.* 2013; **6**(1): 298.
PubMed Abstract | Publisher Full Text | Free Full Text

45. Krajacich BJ, Meyers JI, Alout H, *et al.*: **Analysis of near infrared spectra for age-grading of wild populations of *Anopheles gambiae*.** *Parasit Vectors.* 2017; **10**(1): 552.
PubMed Abstract | Publisher Full Text | Free Full Text

46. Lambert B, Sikulu-Lord MT, Mayagaya VS, *et al.*: **Monitoring the Age of Mosquito Populations Using Near-Infrared Spectroscopy.** *Sci Rep.* 2018; **8**(1): 5274.
PubMed Abstract | Publisher Full Text | Free Full Text

47. Fernandes JN, Dos Santos LMB, Chouin-Carneiro T, *et al.*: **Rapid, noninvasive detection of Zika virus in *Aedes aegypti* mosquitoes by near-infrared spectroscopy.** *Sci Adv.* 2018; **4**(5): eaat0496.
PubMed Abstract | Publisher Full Text | Free Full Text

48. Esperança PM, Blagborough AM, Da DF, *et al.*: **Detection of *Plasmodium berghei* infected *Anopheles stephensi* using near-infrared spectroscopy.** *Parasit Vectors.* 2018; **11**(1): 377.
PubMed Abstract | Publisher Full Text | Free Full Text

49. Lin-Vien D, Colthup NB, Fateley WG, *et al.*: **The Handbook of Infrared and Raman Characteristic Frequencies of Organic Molecules.** (Academic Press Inc.,). 1991.
Reference Source

50. Deng BC, Yun YH, Liang YZ, *et al.*: **A new strategy to prevent over-fitting in partial least squares models based on model population analysis.** *Anal Chim Acta.* 2015; **880**: 32–41.
PubMed Abstract | Publisher Full Text

51. Peiris KH, Drolet BS, Cohnstaedt LW, *et al.*: **Infrared Absorption Characteristics of *Culicoides sonorensis* in Relation to Insect Age.** *Am J Agric Sci Technol.* 2014; **2**(2): 49–61.
Publisher Full Text

52. Khoshmanesh A, Christensen D, Perez-Guaita D, *et al.*: **Screening of *Wolbachia* Endosymbiont Infection in *Aedes aegypti* Mosquitoes Using Attenuated Total Reflection Mid-Infrared Spectroscopy.** *Anal Chem.* 2017; **89**(10): 5285–5293.
PubMed Abstract | Publisher Full Text

53. Hastie T, Tibshirani R, Friedman J: **The Elements of Statistical Learning: Data Mining, Inference, and Prediction.** (Springer). 2009.
Publisher Full Text

54. Babayan SA, Sinclair A, Duprez JS, *et al.*: **Chronic helminth infection burden differentially affects haematopoietic cell development while ageing selectively impairs adaptive responses to infection.** *Sci Rep.* 2018; **8**(1): 3802.
PubMed Abstract | Publisher Full Text | Free Full Text

55. Babayan SA, Liu W, Hamilton G, *et al.*: **The Immune and Non-Immune Pathways That Drive Chronic Gastrointestinal Helminth Burdens in the Wild.** *Front Immunol.* 2018; **9**: 56.
PubMed Abstract | Publisher Full Text | Free Full Text

56. Borchers MR, Chang YM, Proudfoot KL, *et al.*: **Machine-learning-based calving prediction from activity, lying, and ruminating behaviors in dairy cattle.** *J Dairy Sci.* 2017; **100**(7): 5664–5674.
PubMed Abstract | Publisher Full Text

57. Lyimo IN, Haydon DT, Russell TL, *et al.*: **The impact of host species and vector control measures on the fitness of African malaria vectors.** *Proc Biol Sci.* 2013; **280**(1754): 20122823.
PubMed Abstract | Publisher Full Text | Free Full Text

58. Molineaux L, Gramiccia G: **The Garki Project. Research on the epidemiology and control of malaria in the Sudan savanna of West Africa**. World Health Organization, 1980.
Reference Source

59. Bennett S: **Log-Logistic Regression Models for Survival Data.** *Appl Stat.* 1983; **32**(2): 165.
Publisher Full Text

60. Collett D: **Modelling Survival Data in Medical Research**. CRC press, 2003.
Reference Source

61. Bruker Optik GmbH: **Platinum ATR Unit A 225 User Instructions**. 2011.
Reference Source

62. Sellmeier W: **Zur Erklärung der abnormen Farbenfolge im Spectrum einiger Substanzen.** *Ann der Phys und Chemie.* 1871; **219**(6): 272–282.
Publisher Full Text

63. Thomas ME, Tropf WJ: **Optical Properties of Diamond**. In *Proceedings of SPIE* (ed. Klocek, P.) 1994; 144–151.
Publisher Full Text

64. Leertouwer HL, Wilts BD, Stavenga DG: **Refractive index and dispersion of butterfly chitin and bird keratin measured by polarizing interference microscopy.** *Opt Express.* 2011; **19**(24): 24061–6.
PubMed Abstract | Publisher Full Text

65. Dowell FE, Noutcha AE, Michel K: **Short report: The effect of preservation methods on predicting mosquito age by near infrared spectroscopy.** *Am J Trop Med Hyg.* 2011; **85**(6): 1093–6.
PubMed Abstract | Publisher Full Text | Free Full Text

66. Gray EM, Bradley TJ: **Physiology of desiccation resistance in Anopheles gambiae and Anopheles arabiensis.** *Am J Trop Med Hyg.* 2005; **73**(3): 553–9.
PubMed Abstract | Publisher Full Text

67. Sikulu-Lord MT, Milali MP, Henry M, *et al.*: **Near-Infrared Spectroscopy, a Rapid Method for Predicting the Age of Male and Female Wild-Type and *Wolbachia* Infected *Aedes aegypti*.** *PLoS Negl Trop Dis.* 2016; **10**(10): e005040.
PubMed Abstract | Publisher Full Text | Free Full Text

68. Gonzalez Jimenez M, Babayan S, Khazaeli P, *et al.*: **Prediction of malaria mosquito species and population age structure using mid-infrared spectroscopy and supervised machine learning.** University of Glasgow, 2018.
http://www.doi.org/10.5525/gla.researchdata.688

69. Cárdenas G, Cabrera G, Taboada E, *et al.*: **Chitin characterization by SEM, FTIR, XRD, and ^{13}C cross polarization/mass angle spinning NMR.** *J Appl Polym Sci.* 2004; **93**(4): 1876–1885.
Publisher Full Text

70. Yahouédo GA, Chandre F, Rossignol M, *et al.*: **Contributions of cuticle permeability and enzyme detoxification to pyrethroid resistance in the major malaria vector *Anopheles gambiae*.** *Sci Rep.* 2017; **7**(1): 11091.
PubMed Abstract | Publisher Full Text | Free Full Text

71. Suarez E, Nguyen HP, Ortiz IP, *et al.*: **Matrix-assisted laser desorption/ionization-mass spectrometry of cuticular lipid profiles can differentiate sex, age, and mating status of *Anopheles gambiae* mosquitoes.** *Anal Chim Acta.* 2011; **706**(1): 157–163.
PubMed Abstract | Publisher Full Text

72. Wood O, Hanrahan S, Coetzee M, *et al.*: **Cuticle thickening associated with pyrethroid resistance in the major malaria vector *Anopheles funestus*.** *Parasit Vectors.* 2010; **3**(1): 67.
PubMed Abstract | Publisher Full Text | Free Full Text

73. LeCun Y, Bengio Y, Hinton G: **Deep learning.** *Nature.* 2015; **521**(7553): 436–444.
PubMed Abstract | Publisher Full Text

74. Maia MF, Kapulu M, Muthui M, *et al.*: **Detection of *Plasmodium falciparum* infected *Anopheles gambiae* using near-infrared spectroscopy.** *Malar J.* 2019; **18**(1): 85.
PubMed Abstract | Publisher Full Text | Free Full Text

75. Babayan S, Gonzalez M: **SimonAB/Gonzalez-Jimenez_MIRS: First public release (Version v1.0).** *Zenodo.* 2019.
http://www.doi.org/10.5281/zenodo.2609356

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 3

Reviewer Report 18 September 2019

https://doi.org/10.21956/wellcomeopenres.16923.r36486

✔ **Yoosook Lee**
Department of Pathology, Microbiology and Immunology, University of California, Davis, Davis, CA, USA

All concerns were addressed.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Medical entomology, population genetics, malaria vectors, bioinformatics.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 2

Reviewer Report 21 August 2019

https://doi.org/10.21956/wellcomeopenres.16821.r36164

? **Yoosook Lee**
Department of Pathology, Microbiology and Immunology, University of California, Davis, Davis, CA, USA

Thanks for the revision.

- Please include the following method (Joy *et al*. 2012[1]) to be part of the introduction paragraph starting "Given these problems with ovary-based assessment, there has been significant investigation of alternative, molecular based approaches..."

- Figure 3 legend - list the items in incorrect order. Needs to be fixed.
- Figure 8 legend - please provide the wavelength numbers in the legend as in the text.
- Figure 13 - the numbers provided in the confusion matrix (e.g. 82% accuracy in AG classification) doesn't seem to be consistent with the violin plot showing the median less than 80% in both species. Given explanation of the text, it seems like the Figure 13A is still not the final model that provided 82% accuracy. Why not put the results of the best model you found here instead of intermediate result?
- Also put a unit in parenthesis (e.g. %) after prediction accuracy in Figure 13A and Figure 14A and 14C.
- The section of "Predicting mosquito age structure" is problematic. The condition of "increased mortality due to vector control interventions" is not clearly stated. What type of vector control do you have in mind? What's the basis for age-dependent mortality for this? This is also based on a model that is poorly performed in most age groups. I would cut this part and reserve for future studies with improved age prediction models and with some base field data. It is hard to evaluate at current stage.
- This is not necessary condition for indexing but I wonder if the authors would be open to investigate more coercer age groups like in Joy *et al*. 2012[1] - e.g. nonvectors (0–5 days post-emergence), unlikely vectors (6–14 days post-emergence), and potential vectors (15+ days post-emergence) and see if the MIRS would perform better in predicting those age groups (the exact ranges subject to change based on the consensus of the previous reports).
- "We have shown that the variation of MIR spectra over mosquito age can be exploited by a machine-learning algorithm to predict the chronological age with a high degree of accuracy, and ultimately reconstruct population age structures of two important malaria vector species under simulated conditions of changing mortality risk due to vector control". Change to "We have shown that the variation of MIR spectra over mosquito age can be exploited by a machine-learning algorithm to predict the chronological age."
- Also change discussion consistent with the removal of modeling section. For example, change "The ability of this proposed technique to predict the age structure of a population suggests that this approach could constitute an efficient tool for monitoring the efficacy of vector control interventions." to "The model can be used to predict the age structure of field population under various vector control interventions that may impact certain age groups more than others."
- For the discussion paragraph starting "We have shown that MIRS can discriminate between morphologically identical *An. gambiae s.l.* ...." I would add a sentence about how this method could be used as a triage method involving large number (>500-1000) of specimens that needs to be processed in rapid fashion.
- "...increased cuticle thickness that in turn reduces insecticide resistance uptake" - the following reference by Wood *et al*. (2010)[2] seems more appropriate citation for this. Also based on studies by Cook *et al*. (2006)[3] it seems like cuticle thickness would change over time and mature at the age of 10 or above. If true, the unreliability of age determination on the present study may result from the varying degree in speed of maturation at individual level. It may be useful to have some

study to correlate cuticle thickness or kdr genotype and MIRS spectrum at different age groups to determine the maturation of any particular insecticide resistatance trait or cross interaction between kdr genotype, cuticle thickness and insecticide resistance phenotype.

- For journal team, a link to the "Supplementary Note 1" in the Results-Mosquitoes preparation section would be very helpful to navigate the content of the article.

References

1. Joy TK, Jeffrey Gutierrez EH, Ernst K, Walker KR, Carriere Y, Torabi M, Riehle MA: Aging field collected Aedes aegypti to determine their capacity for dengue transmission in the southwestern United States.*PLoS One*. 2012; **7** (10): e46946 PubMed Abstract | Publisher Full Text
2. Wood O, Hanrahan S, Coetzee M, Koekemoer L, Brooke B: Cuticle thickening associated with pyrethroid resistance in the major malaria vector Anopheles funestus.*Parasit Vectors*. 2010; **3**: 67 PubMed Abstract | Publisher Full Text
3. Cook PE, Hugo LE, Iturbe-Ormaetxe I, Williams CR, Chenoweth SF, Ritchie SA, Ryan PA, Kay BH, Blows MW, O'Neill SL: The use of transcriptional profiles to predict adult mosquito age under field conditions.*Proc Natl Acad Sci U S A*. 2006; **103** (48): 18060-5 PubMed Abstract | Publisher Full Text

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Medical entomology, population genetics, malaria vectors, bioinformatics.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 28 Aug 2019

Klaas Wynne, University of Glasgow, Glasgow, UK

Reply to reviewer comments on “Prediction of mosquito species and population age structure using mid-infrared spectroscopy and supervised machine learning [version 2; peer review: 1 approved with reserves]”

Dear Dr. Lee:

Thank you very much for your second report. We really appreciate your valuable comments that have allowed us to improve our manuscript and remove the remaining typos. Please find below our response to your point specific comments:

- Please include the following method (Joy *et al*. 2012) to be part of the introduction paragraph starting "Given these problems with ovary-based assessment, there has been significant investigation of alternative, molecular based approaches..."

Suggestion attended to.

- Figure 3 legend - list the items in incorrect order. Needs to be fixed.

Suggestion attended to.

- Figure 8 legend - please provide the wavelength numbers in the legend as in the text.

Suggestion attended to.

- Figure 13 - the numbers provided in the confusion matrix (e.g. 82% accuracy in AG classification) doesn't seem to be consistent with the violin plot showing the median less

than 80% in both species. Given explanation of the text, it seems like the Figure 13A is still not the final model that provided 82% accuracy. Why not put the results of the best model you found here instead of intermediate result?

- Also put a unit in parenthesis (e.g. %) after prediction accuracy in Figure 13A and Figure 14A and 14C.

We thank the reviewer, but respectfully disagree. The violin plots in Fig 13A describe the performance of multiple models trained on different subsets of the full dataset – they describe a population of model-dataset mappings from which we selected a 'best model' shown in Fig 13B. It is important to show this distribution to be as clear as possible about how the algorithms perform for different subsets of the data and to avoid any doubt about our model selection process. It is also an important figure for supporting the generalisability of our models: regardless of the composition of the training and test sets, the models tend to perform far better than random (red line in Fig 13A). We have clarified the main text of the results and the legend of Figure 13 accordingly.

- The section of "Predicting mosquito age structure" is problematic. The condition of "increased mortality due to vector control interventions" is not clearly stated. What type of vector control do you have in mind? What's the basis for age-dependent mortality for this? This is also based on a model that is poorly performed in most age groups. I would cut this part and reserve for future studies with improved age prediction models and with some base field data. It is hard to evaluate at current stage.

We believe we have already described what type of intervention we have assumed (i.e. an insecticide-treated bed net) as detailed in the method section "*Age-structure modelling*", where indeed this reads: "For the age structure of the populations under a potential intervention regime, we assume that the intervention quadruples the mortality rate of both species from day 3 onwards. This emulates a scenario where mosquitoes encounter an insecticide-treated bednet for the first time on day 3, when they start feeding". To further clarify this, we have now explicitly stated this is a *theoretical* vector control intervention.

We wish to retain this part as this provides a theoretical example of how these predictions and relative modelled age structures perform under different scenarios. We agree that field data will be essential for the empirical evaluation of the technique, which is over the scope of this work, where instead we have evaluated the approach under two theoretical (but plausible) age structure scenarios.

- This is not necessary condition for indexing but I wonder if the authors would be open to investigate more coercer age groups like in Joy *et al*. 2012 - e.g. nonvectors (0–5 days post-emergence), unlikely vectors (6–14 days post-emergence), and potential vectors (15+ days post-emergence) and see if the MIRS would perform better in predicting those age groups (the exact ranges subject to change based on the consensus of the previous reports).

Thank you for this idea. We will take it into account in our next experiments.

- "We have shown that the variation of MIR spectra over mosquito age can be exploited by a machine-learning algorithm to predict the chronological age with a high degree of accuracy, and ultimately reconstruct population age structures of two important malaria vector species under simulated conditions of changing mortality risk due to vector control". Change to "We have shown that the variation of MIR spectra over mosquito age can be exploited by a machine-learning algorithm to predict the chronological age."

Suggestion attended to.

- For the discussion paragraph starting "We have shown that MIRS can discriminate between morphologically identical An. gambiae s.l. ...." I would add a sentence about how this method could be used as a triage method involving large number (>500-1000) of specimens that needs to be processed in rapid fashion.

Thank you very much for this comment. We have added a similar sentence to the text following to your suggestion.

- "...increased cuticle thickness that in turn reduces insecticide resistance uptake" - the following reference by Wood *et al*. (2010) seems more appropriate citation for this. Also based on studies by Cook *et al*. (2006) it seems like cuticle thickness would change over time and mature at the age of 10 or above. If true, the unreliability of age determination on the present study may result from the varying degree in speed of maturation at individual level. It may be useful to have some study to correlate cuticle thickness or kdr genotype and MIRS spectrum at different age groups to determine the maturation of any particular insecticide resistatance trait or cross interaction between kdr genotype, cuticle thickness and insecticide resistance phenotype.

We have changed the reference. Also, we appreciate the interesting suggestion.

- For journal team, a link to the "Supplementary Note 1" in the Results-Mosquitoes preparation section would be very helpful to navigate the content of the article.

Supplementary note 1 was integrated into the methods section. We have corrected this reference.

Competing Interests: No competing interests were disclosed.

Reviewer Report 08 August 2019

https://doi.org/10.21956/wellcomeopenres.16821.r36163

Thomas S. Churcher
MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, Imperial College London, London, UK

The latest version provides greater clarity. The authors did not respond why to the 13 day age group was omitted and the explanation in the text (“The age classes selected were mosquito ages 1, 3, 5, 7, 9, 11, and 15 days, which allowed acceptable per-age accuracy while improving on current binary cut-off of 4 days based on oviposition”), I don’t really understand and am a little nervous about selecting what groupings to present based on their accuracy. Nevertheless, as the authors point out this may not make a difference to the population estimate of age and they provide the raw data which will allow others to explore this should they wish. Overall the manuscript contains a substantial body of work which should form the basis of future research on the technique.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Entomology, epidemiology and statistics.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 02 Sep 2019

Klaas Wynne, University of Glasgow, Glasgow, UK

Thank you.

Competing Interests: No competing interests were disclosed.

Version 1

Reviewer Report 19 June 2019

https://doi.org/10.21956/wellcomeopenres.16586.r35616

?

Thomas S. Churcher
MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, Imperial College London, London, UK

The manuscript presents the first attempt to use mid-infra-red spectroscopy to determine the age and species of Anopheles gambiae s.l mosquitoes. Novel methods for determining the age of the mosquito population are urgently needed and so the new possible technique outlined here is very much welcomed. The paper contains a great deal of information though would benefit from more clarity at times for both the methods and their rationale. Generally, the paper is written with a positive spin and the discussion could be more honest about the frailties of current results (specifically the accuracy) without diminishing the potential of the method. Conversely, the decision to classify mosquitoes into age bands instead of a continuum and the way the results are presented may make the technique look less accurate than it may well be. For example, < 20% of 3 day old *A. gambiae* mosquitoes are classified as being this age, though the confusion matrix appears to show that it is correctly identifying them as being <7 days (though it is hard to see). This level of accuracy could therefore be tolerated depending on the question under investigation, though people might look at the <20% accuracy figure and write it off.

Specific points are detailed below. I am not qualified to comment on technical aspects of spectroscopy.

Major points:

1. Methods. After being killed "*Anopheles gambiae mosquitoes were kept at least one day and An. arabiensis at least two days in the vial to allow them to dry completely*." The two species were therefore, on average, treated differently. Any conclusions about the ability of MIRS to speciate

within the gambiae s.l. complex should therefore be tapered as some of the accuracy maybe because the machine learning was picking up differences in time since death/dryness?

2. Methods. Some spectra with low intensity or high water content were removed by the *Loco Mosquito* 5.0 program automatically. This is available on zenodo yet no criteria for removal were given making it hard to understand what is going on. Did this preferentially exclude mosquitoes of a certain class (you can imagine water content might be related to age)?

3. Methods. Seventeen pre-selected wavelengths were used in the analysis and all other data was excluded. Why this number? It seems strange when there may be considerable information lost. Machine learning methods are designed to deal with this complexity and, though I agree about the dangers of over-fitting, this could be accounted for.

4. Methods. The machine learning methods classified mosquitoes into 7 ages classes. Why was this chosen over a continuous estimation of age? Some justification would be nice as it is different to what is typically done and indeed the authors in the results suggest there was overlap between the bands and training on daily scans (as will happen in the field) made the accuracy worse again. How was this number of classes decided upon? Previous published work has classified mosquitoes as young and old, which would have made the method look much better (it appears from the confusion matrix, it is hard to tell from the shading (consider putting numbers in). Why do those selected not cover the whole range of ages needed? Figure 16A suggests considerable numbers might be older than 15 days which would be important to the average age estimates.

5. Methods “Mosquito ages 1–15 days, taken every two days”. This does not appear to be the case as no data presented from day 13. Why have different boundaries between the oldest two classes? Is the accuracy of predicting 15 age old mosquitoes only better because there are no other classes within 4 days of them when the others have potentially 4 classes?

6. Results. The language of the results sometimes doesn’t come across as very balanced and veers off into opinion. For example, the statement “These results demonstrate that the MIRS signal is strongly indicative of mosquito species and can be used to distinguish between species in a more time and cost efficient method,” not everyone would agree with given the accuracy.

Minor points:

1. Abstract. Species accuracy is given as 82% when it only achieved an average of 76% using the steps outlined in the methods. Dropping 90 of the worst performing models gets you to the better value, but this is only done post hoc. You should either say you are going to do this in the methods or change the headline value in the abstract as it is currently confusing when the reader looks at the abstract and then Figure 13A.

2. Penultimate paragraph page 4. “16 scans were taken at room temperature”. Why was this number done and were they averaged for the spectra used in the analyses?

3. There are lots of methods in the results section. Some of this is repeated and some isn’t, the former could be deleted but the later could be transferred to the methods to aid the reader.

4. The discussion states: *“this method could accurately estimate the proportion of mosquitoes within the older and most epidemiologically-important age classes that responsible for malaria transmission”,* though mosquitoes <15 days of age are likely to be infectious. Suggest revising.

5. The discussion rather grandly says "*To our knowledge, this is the first time that a proposed technique has had the ability to predict the age structure of a population or reconstruct demographic patterns as anticipated from specific vector control interventions*". I'm not sure this is true, as other techniques have showed the ability but the real proof of the pudding comes when moving to the field with the considerable difference between mosquitoes (which the author highlights). Consider revising.

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Partly

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: I am on the same collaborative project as one of the middle authors (Prof Heather Ferguson) though we do not currently work directly together.

Reviewer Expertise: Entomology, epidemiology and statistics.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 30 Jul 2019

Klaas Wynne, University of Glasgow, Glasgow, UK

Reply to reviewer comments on "Prediction of mosquito species and population age structure using mid-infrared spectroscopy and supervised machine learning [version 1; peer review: 1 approved with reservations, 1 not approved]"

Dear Dr. Churcher:

Thank you very much for taking the time to review our manuscript and for your helpful comments, which have enabled us to improve it. We detail point by point below how we addressed each remark.

Methods. After being killed "Anopheles gambiae mosquitoes were kept at least one day and An.

arabiensis at least two days in the vial to allow them to dry completely." The two species were therefore, on average, treated differently. Any conclusions about the ability of MIRS to speciate within the gambiae s.l. complex should therefore be tapered as some of the accuracy maybe because the machine learning was picking up differences in time since death/dryness?

Thank you very much for this comment. In the text we have tried to emphasise the fact that the *An. arabiensis* mosquitoes, because of their physiognomy, usually take one day more than *An. gambiae* to dry with silica gel. However, in our experiments all mosquitoes of both species were dried for more than three days and therefore, no significant differences in humidity that could influence the learning process are expected. We have modified the text of the methods section to clarify this point.

Methods. Some spectra with low intensity or high water content were removed by the Loco Mosquito 5.0 program automatically. This is available on zenodo yet no criteria for removal were given making it hard to understand what is going on. Did this preferentially exclude mosquitoes of a certain class (you can imagine water content might be related to age)?

We have now included in this section how the algorithms of "Loco mosquito" work. Since all the mosquitoes were dry, the algorithm we originally used to detect mosquitoes with high water content was no longer necessary and was removed. We have also corrected this in the methods section.

Methods. Seventeen pre-selected wavelengths were used in the analysis and all other data was excluded. Why this number? It seems strange when there may be considerable information lost. Machine learning methods are designed to deal with this complexity and, though I agree about the dangers of over-fitting, this could be accounted for.

The 17 wavenumbers (13 peak maxima and 4 troughs) were selected to maintain an adequate ratio between inputs and data available, with the aim to collect the maximum information from the spectra in light of their chemical signatures without causing overfitting, a problem that had affected the previous approaches cited in our discussion. Our objective was to prove that this approach works and, as we indicated in the text, we aim to further refine the combination of MIRS and machine learning and adapt it as needed when more field data become available. We have modified the manuscript accordingly.

Methods. The machine learning methods classified mosquitoes into 7 ages classes. Why was this chosen over a continuous estimation of age? Some justification would be nice as it is different to what is typically done and indeed the authors in the results suggest there was overlap between the bands and training on daily scans (as will happen in the field) made the accuracy worse again. How was this number of classes decided upon? Previous published work has classified mosquitoes as young and old, which would have made the method look much better (it appears from the confusion matrix, it is hard to tell from the shading (consider putting numbers in). Why do those selected not cover the whole range of ages needed? Figure 16A suggests considerable numbers might be older than 15 days which would be important to the average age estimates.

Many thanks for this pertinent question. The age classes were selected as a compromise between granularity of the predictions and model performance. We had initially used every day (15 classes), but the models tended to perform poorly for every other age class. As mentioned in the discussion, we hypothesised that this was due to poor differentiation between subsequent days in the MIRS

signals. We expect that more complex models and greater sample sizes should resolve this issue in future studies. However, it is important to note that we do not actually need day-by-day age estimates to reconstruct the age structure of the sampled mosquito population (as we show in Figure 4). This should greatly help translation of this approach to field conditions in due course.

Methods "Mosquito ages 1–15 days, taken every two days". This does not appear to be the case as no data presented from day 13. Why have different boundaries between the oldest two classes? Is the accuracy of predicting 15 age old mosquitoes only better because there are no other classes within 4 days of them when the others have potentially 4 classes?

Addressed.

Results. The language of the results sometimes doesn't come across as very balanced and veers off into opinion. For example, the statement "These results demonstrate that the MIRS signal is strongly indicative of mosquito species and can be used to distinguish between species in a more time and cost efficient method," not everyone would agree with given the accuracy.

We have reviewed results and discussion sections and corrected expressions that appear to veer off into opinion.

Abstract. Species accuracy is given as 82% when it only achieved an average of 76% using the steps outlined in the methods. Dropping 90 of the worst performing models gets you to the better value, but this is only done post hoc. You should either say you are going to do this in the methods or change the headline value in the abstract as it is currently confusing when the reader looks at the abstract and then Figure 13A.

Thank you very much for this comment. We indicate in the methods section that the accuracy of 82.6% is the bagged estimate from the top 10 models. The legend of Figure 13 also explains this.

Penultimate paragraph page 4. "16 scans were taken at room temperature". Why was this number done and were they averaged for the spectra used in the analyses?

We have chosen to average 16 scans because this approach provides noiseless spectra in under a minute. We added more information about this in the text.

There are lots of methods in the results section. Some of this is repeated and some isn't, the former could be deleted but the later could be transferred to the methods to aid the reader.

We are very grateful for this suggestion, but we consider that the results of the experiments carried out to develop our approach from scratch are relevant and, although that they could appear in the methods section, we believe they fit better in the results section given the strong element of methods development.

The discussion states: "this method could accurately estimate the proportion of mosquitoes within the older and most epidemiologically-important age classes that responsible for malaria transmission", though mosquitoes <15 days of age are likely to be infectious. Suggest revising.

We agree that mosquitoes <15 days of age can be infectious, indeed what we suggest here is that the ability of the described approach to determine the whole-age distribution of mosquito

populations (and not only individual age) is important to estimate the proportion that could be epidemiologically-important (which will vary in different settings). We have modified the text to clarify this point.

The discussion rather grandly says "To our knowledge, this is the first time that a proposed technique has had the ability to predict the age structure of a population or reconstruct demographic patterns as anticipated from specific vector control interventions". I'm not sure this is true, as other techniques have showed the ability but the real proof of the pudding comes when moving to the field with the considerable difference between mosquitoes (which the author highlights). Consider revising.

We have now revised this accordingly.

Competing Interests: No competing interests were disclosed.

Reviewer Report 31 May 2019

https://doi.org/10.21956/wellcomeopenres.16586.r35614

Yoosook Lee
Department of Pathology, Microbiology and Immunology, University of California, Davis, Davis, CA, USA

I am deeply troubled by the accuracy of the age prediction (best model was less than 60% - see Figure 3 of this article). That is not much better than flipping the coin. Species prediction is not much better with less than 80% accuracy (Figure 3). Even with bootstrapping (Figure 14), it may accurately call very young (1 day old) and very old (15 days old for *gambiae* and acceptable 80% on 11 day old *arabiensis*) but all other age groups had really poor accuracy (Figure 14). Given the low accuracy, it is hard to imagine the reliability of method as it currently stands. The current paper is written advocating this very low accuracy method. Therefore, I cannot recommend this article to others with good conscience. If the report is written in such way that discusses the limitation of this methods, then I think it is reasonable to publish. There is nothing wrong with trying and thoroughly evaluating the results.

In the hope to improve this article, here are my suggested edits:

In the abstract, title, or keywords there is no mention where this study is conducted. This is significant information because depending on the part of Africa, different members of *Anopheles gambiae* complex exists so the utility of testing done here distinguishing *Anopheles gambiae* from *An. arabiensis* may not transfer well in other regions where *An. gambiae s.s.* and *An. coluzzii* coexists. It is up for further study but the manuscript should be clear on the limited scope of testing done for this study to help readers contemplating using this method.

Abstract "The accuracy of classifying mosquitoes by species was 82.6%" -> Please provide variation or standard deviation of this estimate. Also the Figure 3 data shows all less than 80% so I don't know how

the average could be 82.6%. What about the number for age prediction accuracy? Also state which model was best performing in predictions and the provide correct assignment probability for that in the abstract as author explored multiple models.

“a negligible cost” -> provide an actual estimation in numbers (dollar or euro amount) and what’s included in the cost (equipment, labor, supplies).

“a promising alternative to current mosquito species” -> Many would disagree that 82.6% accuracy in species diagnostic is a promising alternative. I can see the method could be a triage method for selecting specie of interest when dealing with large number of samples for genetic screening.

First paragraph of introduction – because the method introduced here is nothing to do with insecticide resistance, this paragraph containing lengthy intro about the involvement of insecticide resistance seem unnecessary and irrelevant. The second is direct to the point of this paper and good and succinct starting point.

“Currently, *An. gambiae* s.l. species are best distinguished by polymerase chain reaction (PCR) methods [33,34]” -> The citation 33 is no longer valid because it doesn’t distuiguish *An. gambiae s.s.* from *An. coluzzii.* The methods by Favia *et al*. (2001)[1], Fanello *et al*. (2002)[2], and Santolamazazza *et al*. (2008)[3] would be more appropriate for citation here.

“polymerase chain reaction (PCR) methods[33,34], which are time-consuming and expensive”. It is relative to call this method “expensive” as it would cost $2-3 US dollars to screen a single sample as opposed to Lee *et al*. (2015)[4] which cost $5-6 USD for a single sample but collects a wide variety of epidemiologically important traits. Of course, SNP chip or sequencing approach would cost in order of several hundreds per sample. So this molecular assays cited are on the cheap side already. Perhaps more accurate to say “which are time-consuming and still relatively costly, and can thus only be carried out on a subsample of mosquitoes collected during typical entomological surveillance conducted by many agencies in Africa”

Method section mosquito rearing – it is not clear where the mosquito rearing occurred. Is it in Tanzania or Glasgow? Specify the location of *An. arabiensis* Ifakara strain was established.

Method section “Mosquitoes were collected 2 days after a blood meal before egg laying (gravid) and 2 days after an oviposition cup was introduced into the cage (sugar fed).” – This part is confusing.

“mid-infrared spectroscopy-based prediction of mosquito age structure was ….” -> Capitalize the first letter of the sentence.

At the end of 2nd paragraph of introduction: “…the assessment of the impact of these and other control measures”. -> It is not clear what “these and other” control measures refer to. I think it is better to drop the first paragraph and then change “these and other” to “various vector”

“The *An. gambiae* s.l. complex includes several cryptic species that can only be distinguished by molecular analysis” - cryptic species implies two or more species hidden under one species name. I would cite the typical species diagnostic tools only dealing with known “good” species, I would use ‘sibling’ instead of ‘cryptic’.

The introduction paragraph starting with “As in the case of age determination,..” seems out of place. 3rd

and 4th paragraphs talk about age determination, 5th paragraph talks about species determination and 6th paragraph going back to age determination. Some rearrangement of the text necessary to make the story flow better. This paragraph is also very lengthy and makes readers wonder if there is an aging method already available for *Anopheles* species using NIRS. Rather than lengthy and detailed description of what NIRS and why it has problems, it seems the paper is better served by introducing the method of choice and its strength with appropriate citations. To be specific, I would remove the text starting with “while promising, …” till the end of 6th paragraph.

“Here we tested if these limitations can be overcome by shifting the measurement range (25,000-4,000 cm^{-1}) to the mid-infrared region (4,000-400 cm^{-1}), employing an attenuated total reflectance (ATR) device to assess the mosquitoes, and modelling the results with supervised machine learning.”-> In relation to the edits suggested to the previous paragraph, I would change this to “Here, we tested the id-infrared region (4,000-400 cm^{-1}), employing an attenuated total reflectance (ATR) device to assess the mosquitoes which is relatively small.” This suggested edit would make the following phrase in the sentence “Modelling the result with supervised machine learning” awkward. However, the next pagraph introduces why this approach was used so this particular phrase can be removed from this paragraph. Also any reason that the numbers are from large to small?

Method Spectral data acquisition section - “Individual mosquitoes were laid…” -> change it to “dried specimens” to make the experimental process clearer to the readers.

“a globar lamp” -> G should be capital on globar.

“DLaTGS “ -> spell it out like “Deuterated Lanthanum α Alanine doped TriGlycine Sulphate (DLaTGS)”

“KBr beamsplitter” -> spell it out like “Potassium bromide (KBr)”

“Bruker Platinum ATR Unit A225” -> add company name and location of the company
Method Spectral data acquisition “4,000 cm-1” -> “-1” should be superscript.

“Loco Mosquito 5.0” – add a text like “a custom program” to indicate that this program is developed by the authors.

Figure 1. I would prefer if the figure could be self-explanatory as much as possible without having to read legend. Add text or symbol in the figure above each sub-figures what is “correct” and “wrong” way.

Figure 2. It is best if top panel includes an example of “good” reading to differentiate the bad readings. Also I would add text on each panel like “atmospheric intrusion”, “high water content”, and “low intensity”. Pink and red are hard to distinguish. Perhaps dotted lines for pink?

“mod-els” -> “models”

“suit-ed” -> “suited”

“ma-chine” -> “machine”

References

1. Favia G, Lanfrancotti A, Spanos L, Sidén-Kiamos I, Louis C: Molecular characterization of ribosomal DNA polymorphisms discriminating among chromosomal forms of Anopheles gambiae s.s.*Insect Mol Biol*

. 2001; **10** (1): 19-23 PubMed Abstract
2. Fanello C, Santolamazza F, della Torre A: Simultaneous identification of species and molecular forms of the Anopheles gambiae complex by PCR-RFLP.*Med Vet Entomol*. 2002; **16** (4): 461-4 PubMed Abstract
3. Santolamazza F, Mancini E, Simard F, Qi Y, Tu Z, della Torre A: Insertion polymorphisms of SINE200 retrotransposons within speciation islands of Anopheles gambiae molecular forms.*Malar J*. 2008; **7**: 163 PubMed Abstract | Publisher Full Text
4. Lee Y, Weakley AM, Nieman CC, Malvick J, Lanzaro GC: A multi-detection assay for malaria transmitting mosquitoes.*J Vis Exp*. 2015. e52385 PubMed Abstract | Publisher Full Text

Is the rationale for developing the new method (or application) clearly explained?
Partly

Is the description of the method technically sound?
Partly

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
No

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Medical entomology, population genetics, malaria vectors, bioinformatics.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 05 Jun 2019

Klaas Wynne, University of Glasgow, Glasgow, UK

Dear Dr. Lee:

Thank you very much for your report. We appreciate the time you took to read our manuscript and provide constructive feedback. We believe that your major area of concern regarding the manuscript is a result of a misunderstanding about the accuracy of our method that we would now like to clarify.
In the first paragraph of your review, you stated:

I am deeply troubled by the accuracy of the age prediction (best model was less than 60% - see

Figure 3 of this article). That is not much better than flipping the coin. Species prediction is not much better with less than 80% accuracy (Figure 3). Even with bootstrapping (Figure 14), it may accurately call very young (1 day old) and very old (15 days old for gambiae and acceptable 80% on 11 day old arabiensis) but all other age groups had really poor accuracy (Figure 14). Given the low accuracy, it is hard to imagine the reliability of method as it currently stands.

As we indicated in the paper (page 5, version 1), Figure 3 represents a comparison between different algorithms using default (i.e., non-optimised) settings. We do not claim nor suggest that these represent our 'best models', which don't appear in the paper until Figure 13. Indeed, our best models achieved 82.6%, not 60% accuracy.

Because the age model is predicting one of 7 balanced age classes, the null expectation at random is 14.3% accuracy, not a coin flip which is 50%. Consequently, our models need only exceed 14.3% to be better than chance alone. In other words, this is a multi-class rather than binary classification model.

To avoid confusion, we have indicated in figures 13A, 14A, and 14C the accuracy expected under chance alone (red horizontal dashed line) and further clarified the difference between Figures 3 and 13/14 in the text and the caption of Figure 3.

We would like to thank you again for the suggestions to improve our manuscript. This bullet point list shows how we have addressed them.

- *In the abstract, title, or keywords there is no mention where this study is conducted. This is significant information because depending on the part of Africa, different members of Anopheles gambiae complex exists so the utility of testing done here distinguishing Anopheles gambiae from An. arabiensis may not transfer well in other regions where An. gambiae s.s. and An. coluzzii coexists. It is up for further study but the manuscript should be clear on the limited scope of testing done for this study to help readers contemplating using this method.*

We thank the reviewer for this suggestion. To address your concern, we now explicitly indicated in the abstract that the work was developed using laboratory colonies and specified that further study is needed to test this approach on wild populations. Also, in the keywords the two mosquito species are now indicated as Anopheles gambiae s.s. and Anopheles arabiensis, to avoid any ambiguity.

- *Abstract "The accuracy of classifying mosquitoes by species was 82.6%" -> Please provide variation or standard deviation of this estimate.*

The accuracy of 82.6% is the bagged estimate from the top 10 models. The standard deviation on those 10 models is ± 0.002%, but all that says is that the 10 models are all very similar to each other. Therefore, we believe that this number is not very useful and have elected not to include this number.

- *Also the Figure 3 data shows all less than 80% so I don't know how the average could be 82.6%. What about the number for age prediction accuracy? Also state which model was best performing in predictions and the provide correct assignment probability for that in the abstract as author explored multiple models*

Please see response to general points above.

- *"a negligible cost" -> provide an actual estimation in numbers (dollar or euro amount) and what's included in the cost (equipment, labor, supplies).*

Page 17 includes a description of the cost of our method.

- *"a promising alternative to current mosquito species" -> Many would disagree that 82.6% accuracy in species diagnostic is a promising alternative. I can see the method could be a triage method for selecting specie of interest when dealing with large number of samples for genetic screening.*

We believe it is a promising alternative. This paper represents an initial assessment that shows that machine learning can be used to predict traits from the complex infrared spectra of mosquitoes' cuticles and that it is possible to develop a method to survey mosquitos efficiently and with a negligible cost. We believe that this is only the beginning as the evidence presented in the manuscript strongly suggests that this approach is amenable to vast improvements as more field data become available.

- *First paragraph of introduction – because the method introduced here is nothing to do with insecticide resistance, this paragraph containing lengthy intro about the involvement of insecticide resistance seem unnecessary and irrelevant. The second is direct to the point of this paper and good and succinct starting point.*

We believe this first paragraph is important as it introduces the current context (increased insecticide resistance in malaria vector populations) where monitoring mosquito species and age structure is becoming essential to assess the effectiveness of control tool. Without this first paragraph, this context would be lost, so we politely disagree with this point.

- *"Currently, An. gambiae s.l. species are best distinguished by polymerase chain reaction (PCR) methods 33,34" -> The citation 33 is no longer valid because it doesn't distuiguish An. gambiae s.s. from An. coluzzii. The methods by Favia et al. (2001)1, Fanello et al. (2002)2, and Santolamazazza et al. (2008)3 would be more appropriate for citation here.*

Suggestion attended to. Thank you very much for these references.

- *"polymerase chain reaction (PCR) methods33,34, which are time-consuming and expensive". It is relative to call this method "expensive" as it would cost $2-3 US dollars to screen a single sample as opposed to Lee et al. (2015)4 which cost $5-6 USD for a single sample but collects a wide variety of epidemiologically important traits. Of course, SNP chip or sequencing approach would cost in order of several hundreds per sample. So this molecular assays cited are on the cheap side already. Perhaps more accurate to say "which are time-consuming and still relatively costly, and can thus only be carried out on a subsample of mosquitoes collected during typical entomological surveillance conducted by many agencies in Africa"*

Suggestion attended to.

- *Method section mosquito rearing – it is not clear where the mosquito rearing occurred. Is it in Tanzania or Glasgow? Specify the location of An. arabiensis Ifakara strain was established.*

We have now specified that mosquito rearing has occurred at the University of Glasgow; we also indicated that An. arabiensis Ifakara strain was initially established in Tanzania at the Ifakara Health Institute and then also reared at the University of Glasgow.

- *Method section "Mosquitoes were collected 2 days after a blood meal before egg laying (gravid) and 2 days after an oviposition cup was introduced into the cage (sugar fed)." – This part is confusing.*

We have now modified this section to make it clearer. It now reads: "Mosquitoes under three types of physiological conditions were collected, specifically: mosquitoes that had just received a blood meal (blood fed), mosquitoes that had developed eggs as they received a blood meal two days before collection (gravid), and mosquitoes that had laid eggs as they received a blood meal four days before collection and had the chance to lay eggs on an oviposition cup for two consecutive nights (sugar fed)."

- *“mid-infrared spectroscopy-based prediction of mosquito age structure was ….” -> Capitalize the first letter of the sentence.*

Suggestion attended to.

- *At the end of 2nd paragraph of introduction: “…the assessment of the impact of these and other control measures”. -> It is not clear what “these and other” control measures refer to. I think it is better to drop the first paragraph and then change “these and other” to “various vector”.*

We have changed the paragraph to include your suggestion. Thank you very much.

- *“The An. gambiae s.l. complex includes several cryptic species that can only be distinguished by molecular analysis” - cryptic species implies two or more species hidden under one species name. I would cite the typical species diagnostic tools only dealing with known “good” species, I would use ‘sibling’ instead of ‘cryptic’.*

We have now removed the adjective cryptic which can be misleading and used sibling species instead.

- *The introduction paragraph starting with “As in the case of age determination,..” seems out of place. 3rd and 4th paragraphs talk about age determination, 5th paragraph talks about species determination and 6th paragraph going back to age determination. Some rearrangement of the text necessary to make the story flow better. This paragraph is also very lengthy and makes readers wonder if there is an aging method already available for Anopheles species using NIRS. Rather than lengthy and detailed description of what NIRS and why it has problems, it seems the paper is better served by introducing the method of choice and its strength with appropriate citations. To be specific, I would remove the text starting with “while promising, …” till the end of 6th paragraph.*

We have modified the indicated paragraphs following your suggestions to facilitate the reading. However, we have not removed the part about near infrared spectroscopy. We believe that given the similarity at first sight between NIRS techniques and our approach, we should clarify from the introduction why our method is novel and what are its advantages.

- *“Here we tested if these limitations can be overcome by shifting the measurement range (25,000-4,000 cm-1) to the mid-infrared region (4,000-400 cm-1), employing an attenuated total reflectance (ATR) device to assess the mosquitoes, and modelling the results with supervised machine learning.”-> In relation to the edits suggested to the previous paragraph, I would change this to “Here, we tested the id-infrared region (4,000-400 cm-1), employing an attenuated total reflectance (ATR) device to assess the mosquitoes which is relatively small.” This suggested edit would make the following phrase in the sentence “Modelling the result with supervised machine learning” awkward. However, the next pagraph introduces why this approach was used so this particular phrase can be removed from this paragraph.*

Please see above.

- *Also any reason that the numbers are from large to small?*

It is customary among IR spectroscopists to use wavenumbers as energy unit and express it from higher to lower (because it is the reciprocal unit of wavelength, the unit that was traditionally used).

- *Method Spectral data acquisition section - “Individual mosquitoes were laid…” -> change it to “dried specimens” to make the experimental process clearer to the readers.*

Suggestion attended to.

- *“a globar lamp” -> G should be capital on globar.*

Suggestion attended to.

- *“DLaTGS “ -> spell it out like “Deuterated Lanthanum α Alanine doped TriGlycine Sulphate (DLaTGS)”*

Suggestion attended to.

- *"KBr beamsplitter" -> spell it out like "b"*

Suggestion attended to.

- *"Bruker Platinum ATR Unit A225" -> add company name and location of the company*

Suggestion attended to.

- *Method Spectral data acquisition "4,000 cm-1" -> "-1" should be superscript.*

Suggestion attended to.

- *"Loco Mosquito 5.0" – add a text like "a custom program" to indicate that this program is developed by the authors.*

Suggestion attended to.

- *Figure 1. I would prefer if the figure could be self-explanatory as much as possible without having to read legend. Add text or symbol in the figure above each sub-figures what is "correct" and "wrong" way.*

Suggestion attended to. We have included in the figure a green check mark and a red cross mark.

- *Figure 2. It is best if top panel includes an example of "good" reading to differentiate the bad readings. Also I would add text on each panel like "atmospheric intrusion", "high water content", and "low intensity". Pink and red are hard to distinguish. Perhaps dotted lines for pink?*

Suggestion attended to. Labels added and pink line substituted by a black dashed line.

- *"mod-els" -> "models"*

Suggestion attended to.

- *"suit-ed" -> "suited"*

Suggestion attended to.

- *"ma-chine" -> "machine"*

Suggestion attended to. We really appreciate your effort to detect all these typos.

Regarding your concern about the availability of our data expressed in the final questions to the reviewer, we have noticed during the editing of our manuscript the DOI link to our data has changed. We are going to ask the editors to restore the original link (https://doi.org/10.5525/gla.researchdata.688)

Competing Interests: none